Some contributions to the theory and methodology of Markov chain Monte Carlo

Samuel Livingstone

Submitted for the degree of Doctor of Philosophy at the Department of Statistical Science, University College London.

January 13, 2016

I, Samuel Livingstone, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

.....

Abstract

The general theme of this thesis is developing a better understanding of some Markov chain Monte Carlo methods. We review the literature in Chapters 1-4, including a short discussion of geometry in Markov chain Monte Carlo.

In Chapter 5 we consider Langevin diffusions. First, a new class of these are derived in which the volatility is made position-dependent, using tools from stochastic analysis. Second, a complementary derivation is given, here using tools from Riemannian geometry. We hope that this work will help develop understanding of the geometric perspective among statisticians. Such derivations have been attempted previously [108, 43], but solutions were not correct in general. We highlight these issues in detail. In the final part discussion is given on the use of these objects in Markov chain Monte Carlo.

In Chapter 6 we consider a Metropolis–Hastings method with proposal kernel $N(x, hG^{-1}(x))$, where x is the current state. After reviewing instances in the literature, we analyse the ergodicity properties of the resulting Markov chains. In one dimension we find that suitable choice of $G^{-1}(x)$ can change these compared to the Random Walk Metropolis case $N(x, h\Sigma)$, for better or worse. In higher dimensions we show that judicious choice of $G^{-1}(x)$ can produce a geometrically converging chain when probability concentrates on an ever narrower ridge as $|x|$ grows, something which is not true for the Random Walk Metropolis.

In Chapter 7 we discuss stability of Hamiltonian Monte Carlo. For a fixed integration time we establish conditions for irreducibility and geometric ergodicity. Some results are confined to one dimension, and some further to a reference class of distributions. We find that target distributions with tails that are in between Exponential and Gaussian are needed for geometric ergodicity. Next we consider changing integration times, and show that here a geometrically ergodic chain can be constructed when tails are heavier than Exponential.

Acknowledgements

Academically I would like to thank my supervisors, Alexandros Beskos and Mark Girolami. Alex for teaching me Mathematics, building my self-confidence, and making sure I enjoyed my research. Mark for believing in my potential and demanding that I realise it, opening doors, and toughening me up. A special mention is needed for Simon Byrne and Michael Betancourt, in many ways two additional supervisors. Mike for intuition, Simon for rigour, and both for being incredibly sharp, forcing me to raise my game. I am also grateful to Guillaume Bouchard and Cedric Archambeau at Xerox Research Centre Europe for useful discussions, and to Xerox for funding my PhD. Finally to the staff and students in the Department of Statistical Science at UCL. I feel privileged to have studied in the company of so many present and future stars of the field.

Anyone will experience tough times during a three year period of his or her life, and I am no different. Thanks to my brother Tom for showing me what family means when I needed help the most. To Joanna, for making this last year so special. And of course to my mum and dad, for far too many things to mention. This is for you.

Inspirational quotes are common in theses. I include two here. The first helped when I was trying to decide whether I would be doing something 'productive' and 'useful' by embarking on an academic career. The second, for different reasons, inspired Andrey Andreyevich Markov over a century ago to develop the mathematical object on which this thesis is based (see page 34 for more detail).

"Don't ask yourself what the world needs. Ask what makes you come alive, and go do it. Because what the world needs is people who have come alive." — H. Thurman.

"Independence is a necessary condition for the Law of Large numbers." — P.A. Nekrasov.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Monte Carlo methods in Statistics

In some sense Monte Carlo methods mark an about turn for the statistician. We move from treating observed data as realisations of random variables from probabilistic models, to simulating random variables from a model of our choosing. Instead of starting from the data we no longer need any. Put another way, however, standard Monte Carlo is a straightforward application of the two most elementary results in Probability: the Law of Large Numbers and the Central Limit Theorem.

Monte Carlo is a method for estimating intractable integrals. Provided we can write the integral as an expectation of some function f with respect to some probability distribution $\pi(\cdot)$, we simply simulate m independent and identically distributed (iid) observations $X_i \sim \pi(\cdot)$ and note that:

$$\bar{f}_m = \frac{1}{m}\sum_i f(X_i) \to \mathbb{E}_\pi[f(X)]$$

with probability one (see e.g. Section 7.5 of [44]). The natural statistical goal is to understand the properties of the estimator $\bar{f}_m$. But using rules of expectations we can see that if $\mathrm{Var}_\pi[f(X)] = \sigma^2 < \infty$ then $\mathrm{Var}[\bar{f}_m] = \sigma^2/m$, and the Central Limit Theorem gives the celebrated asymptotic

$$\sqrt{m}\left(\bar{f}_m - \mathbb{E}_\pi[f(X)]\right) \xrightarrow{d} N(0,\sigma^2)$$

as $m \to \infty$. We are also able to discuss non-asymptotic results using concentration inequalities.

Perhaps the simplest, Chebyshev's inequality, given by

$$\mathbb{P}\left(\left|\bar{f}_m - \mathbb{E}_\pi[f(X)]\right| > a\right) \leq \frac{\sigma^2}{ma^2}, \;\; a > 0,$$

allows us to construct non-asymptotic confidence bounds on estimation error. In many specific cases much sharper bounds exists [16].

The drawback of Monte Carlo in its simplest form is the need to generate independent samples from $\pi(\cdot)$. There are many scenarios in which this is not feasible. We focus in the next section on the most prominent case among statisticians, though certainly not the only one worthy of note (see e.g. [41, 65]).

1.1.1 Bayesian inference

Philosophical debates on the correct approach to inferring unknown quantities have raged for many years, and will doubtless continue (see e.g. Chapter 1 of [114]). At least two methods rely on constructing some kind of probabilistic model (the *likelihood*) for some data y, which depends on a set of parameters θ. In the Bayesian approach, the current state of understanding for θ before observing y is then encoded through a probability distribution, known as the *prior*, with density $\pi_0(\theta)$.[1] Using only the prior and likelihood term $f(y|\theta)$, we are then able to establish a posterior state of knowledge for θ using Bayes' theorem

$$\pi(\theta|y) \propto f(y|\theta)\pi_0(\theta). \tag{1.1}$$

Constructing the posterior distribution up to a constant of proportionality is therefore trivial. However, extracting relevant information from $\pi(\theta|y)$ (such as posterior means, marginal densities and quantiles for parameters of interest) relies on taking expectations with respect to it. As referenced in the previous section, since we only know $\pi(\theta|y)$ up to a constant, ordinary Monte Carlo is typically no longer an option.

1.1.2 Monte Carlo using Markov chains

When independent samples cannot be directly generated there are a number of approaches, collectively termed 're-sampling', in which data are generated from some *candidate* distribution $q(\cdot)$, and

[1] Of course, both discrete and continuous parameters can be represented through a prior, of both finite and infinite dimension, but here we focus on the finite dimensional continuous case for ease of exposition.

an estimator is constructed from these samples for expectations with respect to $\pi(\cdot)$. In 'rejection' sampling, for example, some of the draws from $q(\cdot)$ are discarded, so that what is left is a representative sample from $\pi(\cdot)$. In 'importance' sampling, each draw is weighted according to its importance in inferring quantities from $\pi(\cdot)$.

An approach which has proven fruitful in practice is to simulate a Markov chain with limiting distribution $\pi(\cdot)$. We are then considering the estimator

$$\tilde{f}_m = \frac{1}{m}\sum_i f(X_i), \;\; X_i \sim P^i(x_0, \cdot),$$

where $P^i(x_0, \cdot)$ is the i step transition kernel. Note two things here:

1. The marginal distribution of each X_i is not $\pi(\cdot)$
2. The random variables X_i and X_{i+1} are not independent of each other.

It is somewhat surprising, therefore, that under very mild conditions on the chain

$$\tilde{f}_m \to \mathbb{E}_\pi[f(X)]$$

as $m \to \infty$, with probability one. If we can show in addition that the sequence of marginal distributions $P^i(x_0, \cdot)$ for each X_i converges to $\pi(\cdot)$ at a certain rate, then we can also rely on a Central Limit Theorem [56] result

$$\sqrt{m}\left(\tilde{f}_m - \mathbb{E}_\pi[f(X)]\right) \xrightarrow{d} N(0, v(P, f)) \tag{1.2}$$

for the estimator $\tilde{f}_m$ (note that certain restrictions again must be placed on f, which we discuss later). In some cases we can also appeal to non-asymptotic bounds (we also discuss this in more detail in Chapter 2).

1.1.3 The need for rigorous understanding

Several different Markov chain Monte Carlo (MCMC) algorithms exist today in the Statistics, Mathematics, Physics and Computer Science literature [18], so for a given problem there are several different MCMC estimators to choose from. It is vitally important for practitioners to understand which method to use in a given scenario. Understanding for what forms of $\pi(\cdot)$ a version of (1.2) holds, and the corresponding asymptotic variance $v(P, f)$ for each algorithm, gives a principled way

to make such a choice. Perhaps more disturbingly, without establishing the necessary convergence properties of the Markov chain, we have very little guarantees on the quality of the estimator $\tilde{f}_m$.

In modern applied Statistics, MCMC is ubiquitous [29]. Ensuring that the optimal methods are being used, and understanding the strengths and weaknesses of each, is clearly necessary to ensure that the right insights are being drawn from empirical research. The goal of this thesis is to make a contribution towards this end.

1.1.4 Thesis outline

In Chapters 2, 3 and 4, we review the field of Markov chain Monte Carlo. We begin with stochastic simulation and its use in statistical inference, to motivate why methods based on Markov chains have become popular. We then give a detailed review of Markov processes, mainly (but not exclusively) in discrete time. After this we introduce some common Markov chain Monte Carlo methods, discussing the strengths and weaknesses of each. A highlight of this chapter is a review of geometric concepts in Markov chain Monte Carlo, in Section 4.3, which is based on the author's own published work in [71].

In the rest of the thesis we present some original contributions. Parts of Chapter 5 are based on two published works [130] and [71]. Chapter 6 is based on the submitted work [70]. Chapter 7 is motivated by the submitted work [11], but is mostly more recent work which is currently in preparation. Chapter 8 contains a short summary of the thesis contributions along with some possible avenues for further research.

1.1.5 Notational conventions

We use $\{X_t\}_{t\geq 0}$ to denote the process $\{X_0, X_1, X_2, ...\}$ that evolves in discrete time, and $(X_t)_{t\geq 0}$ for the continuous-time variant $\{X_t : t \in \mathbb{R}_{\geq 0}\}$. Throughout if $\pi(\cdot)$ denotes a probability measure then $\pi(x)$ will be the corresponding density with respect to Lebesgue measure on the measurable space $(\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n))$, with $\mathscr{B}(\mathbb{R}^n)$ the Borel σ-algebra (this is discussed more thoroughly in Section 3.1). We denote by $\mathbf{X}$ the state-space for a Markov chain, and unless otherwise stated this can also be assumed to be $\mathbb{R}^n$. In the discrete time case, we sometimes refer to the m-step transition probabilities for a Markov chain. Most text books refer to n-step transition probabilities, but we reserve n for the dimension of the state space of a random variable. The letter d is sometimes used

for this purpose, but we use d for distance metrics. For an element $x \in \mathbb{R}^n$, we write $|x| = \sqrt{\sum_i x_i^2}$ to denote the Euclidean norm. For a set A, $|A|$ denotes its cardinality, the number of elements in A. Three commonly used measures are $\delta_a(\cdot)$, the Dirac measure at a, for which $\delta_a(A) = 1$ if $a \in A$ and 0 otherwise, $\mu^G(\cdot)$, the standard Gaussian measure on $\mathbb{R}^n$, and $\mu^L(\cdot)$, the standard Lebesgue (or *length*) measure on $\mathbb{R}^n$ (see e.g. Sections 2.3-2.4 of [20]). In the context of Markov processes, we sometimes use the conditional probability notation $\mathbb{P}_x[\cdot] := \mathbb{P}[\cdot|X_0 = x]$ and $\mathbb{E}_x[\cdot] := \mathbb{E}[\cdot|X_0 = x]$. This will mean we are technically conditioning on the null set '$X_0 = x$'. We discuss this issue in Section 3.1.

Chapter 2

Stochastic simulation methods

Since at least the 1970s [38], simulation-based methods have been exploited for a variety of goals in Statistics. Indeed some argue that they have revolutionised the field [29]. The premise that realisations of random variables can be 'simulated' with high precision has allowed more complex statistical models to be developed, for a variety of reasons. In the classical framework, hypotheses can be constructed based on test statistics which no longer need to follow an asymptotic distribution which can be derived analytically, we can simply simulate data under the null hypothesis and approximate the distribution with a histogram. Similarly confidence intervals in linear models need no longer rely on the assumption of normality of errors, thanks to bootstrapping techniques [38]. These are but two examples of many. We focus here on the *Monte Carlo* method, which originated in the Physics literature (e.g. [79]) but has found many useful applications in Statistics [99], for reasons we now discuss.

2.1 Intractable integrals & Monte Carlo

The problem of interest is evaluating intractable integrals, specifically those that can be written as expectations

$$\mathbb{E}_{\pi}[f(X)] = \int f(x)\pi(dx). \tag{2.1}$$

Such integrals arise often in Statistics. A prominent example is an intractable likelihood

$$L(\theta;y) = \int f_\theta(y|x) f_\theta(x) dx,$$

where y represents some observed data, x some unobserved data and θ a parameter of interest. Often conditional on knowing x the likelihood takes a straightforward form, but the marginal likelihood given only y is much more complex. Another, and perhaps the most common example is Bayesian inference, where information about θ is encoded in the *posterior* distribution, and posterior expectations and quantiles must be computed by integration.

2.1.1 Numerical methods

One approach to the problem is to compute the integral numerically. Perhaps the simplest method is an approximation by a collection of rectangles. If we assume that the region of integration is some bounded interval $[a,b]$, then we divide this region into $a = a_1 < a_2 < ... < a_m = b$, where for simplicity we assume the a_i are equidistant with $\triangle a = a_{i+1} - a_i$. We approximate the integral with the sum

$$S_m = \sum_{i=1}^{m-1} f(a_i) \triangle a. \tag{2.2}$$

Clearly as m grows the approximation becomes more accurate.[1] The problem with such grid-based numerical methods is scaling with dimension. The sum above involves $m-1$ terms. In the n-dimensional case, where we are interested in

$$\int f(x_1,...,x_n) dx_1 ... dx_n,$$

we must compute

$$S_m^n = \sum_{i_1=1}^{m-1} ... \sum_{i_n=1}^{m-1} f(a_{i_1},...,a_{i_{m-1}}) (\triangle a)^n,$$

where now the sum S_m^n involves $(m-1)^n$ terms. In short, such grid-based methods typically scale exponentially with dimension. Of added concern to the statistician, we often cannot assess statistical properties of the resulting 'estimates' for the integral, such as bias and efficiency.

[1] Provided that f is suitably well-behaved.

2.1.2 Monte Carlo and random number generation

In ordinary Monte Carlo we simply simulate data from $\pi(\cdot)$ and compute the 'sample average' estimator

$$\hat{f}_m = \frac{1}{m}\sum_{i=1}^{m} f(X_i), \;\; X_i \sim \pi(\cdot).$$

As stated previously, Laws of Large Numbers, Central Limit Theorems and concentration inequalities provide a range of tools with which to assess the quality of this estimator. The method is also (in some sense) dimension-independent: we can see that $\hat{f}_m$ is a sum of m terms, regardless of the dimension of X.[2] A key difference between this and grid-based approaches is that effort is concentrated here on relevant parts of the space, as more samples will be generated in areas which are more likely under $\pi(\cdot)$.

Truly random numbers can be straightforward to generate in reality, by simply rolling a die or flipping a coin. Producing these in large quantities, however, would be time consuming. In the vast majority of cases, those working with stochastic simulation tools instead use *pseudorandom* numbers, which are deterministic sequences produced by a computer, designed so that they are *statistically random*, in the sense that a given sequence is indistinguishable from one that would have been produced from the desired distribution, according to some standard hypothesis tests. Generators are usually designed to produce $U[0,1]$ random variables. A simple eample is the *linear congruence generator*, where $\{u_1, u_2, ...\}$ is determined by the recursion

$$u_{n+1} = (au_n + b) \text{mod} M,$$

where a and M are large coprime integers. See [98] for more detail on the properties of such methods, and recommended choices for a, b and M.[3]

To draw samples from other distributions, usually $U[0,1]$ sequences are first generated and then transformed. This is straightforward to do in many cases because the $U[0,1]$ cumulative distribution function is the identity, i.e. $\mathbb{P}[U < u] = u$. If X follows a distribution such that $\mathbb{P}[X \leq x] = F_X(x)$, then

$$\mathbb{P}[X \leq x] = F_X(x) = \mathbb{P}[U \leq F_X(x)] = \mathbb{P}[F_X^{-1}(U) \leq x],$$

[2] In reality this depends on the function being estimated. If we consider the specific case $f(x) = \prod_{j=1}^{n} x_j$, with $X = (X_1, ..., X_n)$ comprised of iid zero mean components and $\text{Var}[X_j] = \sigma^2$, then $\text{Var}[\hat{f}_m] = \sigma^{2n}/m$, which grows exponentially with dimension.

[3] We leave the discussion on drawing samples of continuous random variables to Section 3.1.

meaning provided F_X is invertible then $F_X^{-1}(U)$ follows the desired distribution.[4] is This method is known as the *probability integral transform* (see Section 2.1.2 of [99]).

Example 2.1. *To generate a sample from an Exponential distribution with rate parameter 1, where $F_X(x) = 1 - e^{-x}$, we generate $u \sim U[0,1]$, and set $x = -\log(1-u)$.*

One example where F_X cannot be inverted analytically is the Gaussian distribution. A simple approach to generating two $N(0,1)$ random variables is using the *Box–Muller* method [17]. The logic is that if $(X,Y) \sim N(0, I_{2\times 2})$ then

$$R^2 = X^2 + Y^2 \sim \chi_2^2, \quad \text{and} \quad \theta = \arctan(Y/X) \sim U[0, 2\pi].$$

Using this polar coordinate transform, we can generate $R^2 = -2\log U_1$ and $\theta = 2\pi U_2$ using the probability integral transform from uniform samples U_1 and U_2, and then recover X and Y. Combining steps gives

$$X = \sqrt{-2\log U_1}\cos(2\pi U_2), \quad Y = \sqrt{-2\log U_1}\sin(2\pi U_2).$$

An n-dimensional Gaussian random variable $X \sim N(\mu, \Sigma)$ can be generated by setting

$$X = \mu + (\sqrt{\Sigma})Z,$$

where $Z \sim N(0, I_{n\times n})$, and $\sqrt{\Sigma}$ is a matrix such that $\sqrt{\Sigma}(\sqrt{\Sigma})^T = \Sigma$, which can be found using (for example) a Cholesky decomposition of Σ [126]. It should be noted that this process is *not* dimension-independent, as the decomposition is $O(n^3)$. However, (anecdotally) such random number generation does not appear to be much of a computational bottleneck in practice for our needs, so we do not discuss it further.

We will be concerned with the case where the *density* of interest (and hence F_X) is only known up to a proportionality constant. Often in the Bayesian context this is the case, as shown in (1.1). Here direct simulation from $\pi(\cdot)$ using appropriate transformations is often not possible.

2.1.3 Re-sampling, indirect Monte Carlo

Often in the above scenario we can still draw samples from $\pi(\cdot)$ using a two stage process:

[4] In the case where X is discrete then the *generalised* inverse can be used, see Chapter 2 of [99].

1. Draw samples from some candidate distribution $q(\cdot)$

2. Modify them in such a way that integrals with respect to $\pi(\cdot)$ can be estimated

Two popular methods are *rejection sampling* and *importance sampling*.

In the first, stage two involves either keeping or discarding each $X \sim q(\cdot)$ with some probability

$$\mathbb{P}[\text{Accept sample } x] = \frac{\pi_u(x)}{Mq(x)},$$

where M is chosen such that this ratio is at most one, and $\pi_u(x)$ represents the *unnormalised* version of $\pi(x)$. It is straightforward to see that

$$\mathbb{P}[X \in A | X \text{accepted}] = \frac{\mathbb{P}[X \in A, X \text{accepted}]}{\mathbb{P}[X \text{accepted}]} = \frac{\int_A q(x) \frac{\pi_u(x)}{Mq(x)} dx}{\int q(x) \frac{\pi_u(x)}{Mq(x)} dx} = \frac{\int_A \pi_u(x) dx}{\int \pi_u(x) dx} = \pi(A),$$

meaning the rejection method produces independent samples from $\pi(\cdot)$. Efficiency is dictated by how regularly samples from $q(\cdot)$ are accepted, which requires this candidate distribution to be chosen such that it is 'similar' to $\pi(\cdot)$ in some sense.

In the basic importance sampling scheme, stage two involves replacing (2.2) with the estimator

$$\bar{f}_m = \sum_{i=1}^{m} f(X_i) \frac{\pi(X_i)}{q(X_i)}. \tag{2.3}$$

Trivially,

$$\int f(x) \frac{\pi(x)}{q(x)} q(x) dx = \int f(x) \pi(x) dx = \mathbb{E}_\pi[f(X)].$$

The ratios $\pi(x)/q(x)$ are known as 'importance weights'. For many functions of interest[5] the method is most effective when each of these weights is as close to one as possible (see Chapter 3 of [99]). Of course, in (2.3) $\pi(x)$ needs to be known exactly, so in the case where only $\pi_u(x)$ is known the modified estimator

$$\check{f}_m = \frac{\sum_{i=1}^m f(X_i) w(X_i)}{\sum_{i=1}^m w(X_i)}, \quad w(X_i) = \frac{\pi_u(X_i)}{q(X_i)} \tag{2.4}$$

is used, which is derived from the expression

$$\mathbb{E}_\pi[f(x)] = \frac{\int f(x) \frac{\pi_u(x)}{q(x)} q(x) dx}{\int \pi_u(x) dx} = \frac{\int f(x) \frac{\pi_u(x)}{q(x)} q(x) dx}{\int \frac{\pi_u(x)}{q(x)} q(x) dx}.$$

For finite m (2.4) has some bias, but is often in fact more efficient than (2.3) as shown in Section 3.3.2 of [99]. It is also important to choose $q(\cdot)$ so that the ratio $\pi(x)/q(x) \to 0$ as $|x| \to \infty$, as this

[5] A notable exception here is function that concentrate on the tails of a distribution, see Chapter 3 of [99] for more detail.

also has an impact on the variance of estimators (2.3) and (2.4), again as discussed in Section 3.3 of [99].

The problem with these methods is again scaling with dimension. Both rely on choosing a candidate distribution $q(\cdot)$ which approximates $\pi(\cdot)$ *globally* in some sense. Often when n is large we have limited knowledge of $\pi(\cdot)$ making a good choice of $q(\cdot)$ an extremely challenging task. For the rejection method, the result will be that typically very few samples are accepted. In the importance sampling case, the variation in importance weights is often very large, meaning the estimator (2.4) is extremely inefficient. Examples illustrating these difficulties are described in detail in [72].

2.2 Markov chain Monte Carlo

We have already introduced the idea of Monte Carlo using Markov chains in the introduction. This section simply serves to motivate Markov chain Monte Carlo (MCMC) further. The key point of note is that the re-sampling methods discussed in the previous section can fail because of the need to understand what $\pi(\cdot)$ looks like globally. Markov chains, on the other hand, can be constructed in such a way that they explore the space *locally*. The question is modified from how to draw a sample which is 'likely under $\pi(\cdot)$' to one of where to move next given the current location in the chain. Evaluating a proposed move y given the current position x can be done through the ratio $\pi(y)/\pi(x) = \pi_u(y)/\pi_u(x)$, which *directly* assesses whether the chain will be moving in a direction which is more or less likely under $\pi(\cdot)$. As a result, MCMC methods can produce estimators for intractable integrals which scale much more favourably with dimension than either numerical or re-sampling counterparts.

In the next section we give a thorough review of the underlying mathematics of Markov chains, which will be exploited to develop new results in this work. We focus mainly on the discrete time case, but also review some continuous-time processes which will be used later.

Chapter 3

Markov chains

A course on Markov chains which unfold on a finite state space is a typical module on any undergraduate Mathematics and Statistics degree. Moving to the case where each X_t is defined on an uncountable space requires some understanding of measure-theoretic probability, and hence a good deal more subtlety. For this reason, while we deal with the general case here, we refer to the finite/countable case periodically to aid intuition for some concepts.

3.1 A note on real numbers & measure theory

Modelling real numbers using Probability theory is both intuitive and extremely puzzling. It is very natural when confronted with a collection of data points $12.345, 18.421, 34.564...$ to consider them as 'continuous'. However, to place a probability distribution over the entire real number line one must set the probability of any specific outcome in $\mathbb{R}$ to zero.

Proposition 3.1. *If $(x_i)_{i\in I}$ is an uncountably large collection of real numbers with each $x_i \geq 0$ such that $\sum_i x_i < \infty$, then $x_i = 0$ for all but at most countably many $i \in I$. Setting each x_i to be the probability of outcome $i \in I$ implies that only a countable number of these can be assigned a non-zero probability.*

Proof: Suppose the sum is finite, so $\sum_i x_i = M < \infty$. We show that $I_{>0} = \{i \in I : x_i > 0\}$ must be

countable. First consider the set S_n of all i for which $x_i > 1/n$. We have

$$M \geq \sum_{i \in S_n} x_i \geq \frac{1}{n}|S_n|,$$

where $|S_n|$ denotes the number of elements in S_n. So S_n can have at most Mn elements, and hence is finite. From here we simply note that $I_{>0} = \cup_{n \in \mathbb{N}} S_n$, which is a countable union of finite sets, and hence is countable. ■

Since we require probabilities to be positive and sum to one, this makes life difficult. The philosophical conundrum of how we uncovered the data we have is resolved by taking measurement precision into account: our data are not in fact real numbers, but each is the set of all real numbers which are equivalent up to a certain decimal place. Our model $\mathbb{R}$ is only ever an approximation to reality, but it will be arbitrarily good for arbitrarily high accuracy of measurements.

Mathematically we first deal with the problem by using densities in place of probability mass functions, using the physical intuition $mass = density \times volume$ to compute probabilities with integrals. However, this makes it hard to have a unified treatment of the theory. For example, one cannot define a random variable which can either take the value 0 (with positive probability) or any positive real number. Additionally the Riemann integral learned during undergraduate Mathematics can be undone by some simple probabilistic questions: for example, we cannot compute the probability that a randomly selected number between 0 and 1 will be rational, as the upper and lower Riemann sums do not converge to each other but remain at 1 and 0 respectively for any partition of $[0,1]$.

The language of measures relieves both of these problems. A measure is simply a function whose argument is a set, that assigns a number to that set. In addition, a measure $\mu(\cdot)$ must be *countably additive*, meaning for any countable collection of sets $\{E_i\}$ with $E_i \cap E_j = \emptyset$ for all i, j we have

$$\mu\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} \mu(E_i). \tag{3.1}$$

Measures are natural to the probabilist as the probability of an event, or set of possible outcomes. Countable additivity is also intuitive, and is in fact a probability axiom. Using this language we can construct probability measures for both discrete and continuous random variables as well as mixtures. For example, a random variable which takes the value 0 with probability $1/2$ or else a uniformly chosen number between 1 and 5 has probability measure

$$\pi(\cdot) = \frac{1}{2}\delta_0(\cdot) + \frac{1}{2}\tilde{\mu}^L(\cdot),$$

where $\tilde{\mu}^L(A) = \mu^L(A \cap [1,5])/4$. Similarly, the Lebesgue integral trivially asserts that the number of elements of $[0,1]$ for which $\mathbb{1}_{x\in\mathbb{Q}}$ is non-zero is a countable collection of points and hence has Lebesgue measure zero (by countable additivity), so there is no chance that a randomly chosen number between 0 and 1 is rational.

Unfortunately, complications arise from Lebesgue's careful study of measure. It can be shown (assuming the axiom of choice) that sets exist for which (3.1) does not hold (see e.g. the Appendix on page 301 of [20]). So in order to discuss Probability rigorously, we must specifiy a measurable space $(\mathbf{X}, \mathscr{B})$, the set of possible outcomes $\mathbf{X}$ combined with a collection of subsets of $\mathbf{X}$ (termed the σ-algebra) for which (3.1) holds, and restrict our analysis to these sets. Probabilistically, we can think of these as the sample space and event space. In practice, we are unlikely to ever need to worry about sets which are not in $\mathscr{B}$. However, for consistency in this thesis we will always work on the abstract probability space $(\Omega, \mathscr{B}(\Omega), \mathbb{P}(\cdot))$, representing the 'state of the universe', and consider random variables as maps $X(\omega)$ from $\omega \in \Omega$ to some $(\mathbf{X}, \mathscr{B})$, with $\mathbf{X}$ a complete, separable metric space (e.g. [112]), and when stated with an induced probability measure $\pi(\cdot)$.[1]

3.1.1 Conditional probability

One particular challenge of continuous random variables is conditioning. A basic course in Probability will teach that

$$\mathbb{P}[X \in A | Y \in B] = \frac{\mathbb{P}[X \in A, Y \in B]}{\mathbb{P}[Y \in B]}, \tag{3.2}$$

where A and B are events and X and Y random variables, and the logic is undeniable from simply drawing a Venn diagram. However, if $\mathbb{P}[Y \in B] = 0$ then (3.2) does not actually make sense. When discussing Markov chains we will often be in this scenario, by attempting to condition on the current state in the chain having a specific value when defining the distribution for the next state.

We resolve this in practice by using (3.2) on non-null sets,[2] and defining a *regular* conditional probability measure for conditioning on specific points in the state-space. Loosely, the idea is that two random variables $X|Y_1$ and $X|Y_2$ will agree *almost surely* provided i) they agree on non-null sets under Y_1 and Y_2, and ii) Y_1 and Y_2 have the same null sets (see e.g. Section 6.5.3 of [20]). So on these null sets we can actually choose from several *versions* of the conditional distribution. A regular

[1] Those unfamiliar with measures will not lose too much by simply replacing $\int \pi(dx)$ with $\int \pi(x)dx$ where appropriate and noting that $\int f(x)\delta_a(dx) = f(a)$.

[2] A null set is simply a set of zero probability under the chosen distribution.

conditonal distribution is just a way of picking a particular version. In the Markov chains case, we will call this version a *transition kernel*, defined in the next section. However, for ease of exposition, we will still sometimes formally write $\mathbb{P}[Y \in A|X = x]$ to denote a conditional probability, as well as occasionally $\mathbb{P}_x[X_t \in A] := \mathbb{P}[X_t \in A|X_0 = x]$. For a much more rigorous and detailed discussion of conditional probability, see [59].

3.2 Markov chains in discrete time

A stochastic process $\{X_t\}_{t\geq 0}$ with each X_i defined on $(\mathbf{X}, \mathscr{B})$ is called a *Markov chain* if

$$\mathbb{P}[X_i \in A|X_{i-1} = x_{i-1}, ..., X_0 = x_0] = \mathbb{P}[X_i \in A|X_{i-1} = x_{i-1}]. \tag{3.3}$$

If in addition the distribution of $X_i|X_{i-1}$ does not depend on i, we call the chain *time-homogeneous* (we only consider chains of this form here). Given (3.3), we can completely characterise $\{X_t\}_{t\geq 0}$ through an *initial* distribution $\mu(\cdot)$ for X_0, and a set of conditional distributions $\mathbb{P}[X_1 \in A|X_0 = x_0]$ for each $x_0 \in \mathbf{X}$.[3] For the latter we use a *transition kernel*

$$P : \mathbf{X} \times \mathscr{B} \to [0, 1].$$

For any fixed $x_0 \in \mathbb{R}$, $P(x_0, \cdot)$ defines a distribution over $(\mathbf{X}, \mathscr{B})$, and for any $A \in \mathscr{B}$, $P(\cdot, A)$ is measurable. Intuitively, P defines a map from points to distributions in $\mathbf{X}$. Similarly, we can define the m-step transition kernel as

$$P^m(x_0, A) = \mathbb{P}[X_m \in A|X_0 = x_0],$$

which we can find recursively through the calculation

$$P^m(x_0, A) = \int P^{m-1}(y, A)P(x, dy).$$

Example 3.2. *A simple stationary Gaussian AR(1) process defined recursively as* $X_{i+1} = \rho X_i + \sqrt{1-\rho^2}\varepsilon_i$, $\varepsilon_i \sim N(0, 1)$ *can be written as a Markov chain with initial distribution* $\mu^G(\cdot)$ *and transition kernel* $P(x, \cdot)$ *defined to be* $N(\rho x, 1 - \rho^2)$*, for any* $\rho \in (-1, 1)$.

[3] The existence of a stochastic process defined via these objects is a straightforward consequence of Kolmogorov's consistency theorem in most cases. See Chapter 3 of [81] for more detail.

In the case where $|\mathbf{X}| = n < \infty$, the transition kernel is simply an $n \times n$ matrix P, and the distribution for any X_i is an n-dimensional row vector ν with $\nu_i \geq 0$ and $\sum_i \nu_i = 1$. The marginal distribution ν' for X_{i+1} can then be written $\nu' = \nu P$. When P is of this form, we can elegantly represent the relationship between transition probabilities through the *Chapman-Kolmogorov* equations

$$P^{m+n} = P^m P^n. \tag{3.4}$$

We must first define the operator P in the general case in order to express things in the same way. We do this through its action on probability measures (to the left) as

$$\nu P(A) = \int P(x,A)\nu(dx), \tag{3.5}$$

with which we can define the marginal distribution of X_1 in the instance $X_0 \sim \nu(\cdot)$, and on functions (to the right) as

$$Pf(x) = \int f(y)P(x,dy), \tag{3.6}$$

which gives the conditional expectation of $f(X_1)$ given that $X_0 = x$. With (3.5) we can write $\nu'(\cdot) = \nu P(\cdot)$ as in the finite case, while (3.6) means we can write (3.4) in the general case. Note that (3.4) is in fact the semi-group property for the family of linear operators $\{P^t\}_{t\geq 0}$ [88].

We are interested in Markov chains as a means to approximate expectations under some distribution from which we cannot simulate directly. The reason that many Markov chains present an avenue to do this is by a property known as *ergodicity*. Informally, for some specific forms of P, there exists a unique distribution $\pi(\cdot)$ for which in some sense

$$\mu P^m(\cdot) \to \pi(\cdot) \tag{3.7}$$

as $m \to \infty$, for any choice of $\mu(\cdot)$. Thus, direct simulation from $\pi(\cdot)$ is no longer required in order to draw samples with distribution 'arbitrarily close' to $\pi(\cdot)$, in a sense that we will make rigorous later. In this instance, we also have a version of the Strong Law of Large Numbers, again informally stated here as

$$\frac{1}{m}\sum_{i=1}^{m} f(X_i) \to \mathbb{E}_\pi[f(X)], \;\; X_i \sim \mu P^i(\cdot). \tag{3.8}$$

In the following sections we make these ideas rigorous, discussing the requirements on P which result in the Markov chain being 'ergodic', and how these translate into guarantees on estimators. To do this, we must first understand long-time behaviour. This can be an arduous task, as the theory is rich. Here we summarise some notions from countable state space chains, before moving on to the

general case. We introduce the countable case first as many concepts seem natural there, and in the general case ideas have often been adapted from those for countable chains. Geometric ergodicity is also most naturally motivated from the countable state space theory. Loosely, in the following sections we will characterise the sets $A \in \mathscr{B}$ which will be visited by the chain, and then those that will be visited *infinitely often*. Any limiting distibution $\pi(\cdot)$ will only have $\pi(A) > 0$ if this is the case, as otherwise $\sum P^m(x,A)/m$ will converge to zero as $m \to \infty$. We then establish conditions for convergence to a unique limit $\pi(\cdot)$, rates of convergence, and corresponding limit theorems for estimators taken from chains.

Historical Note. Interestingly, Markov chains were in fact specifically designed as a process for which (3.8) could be established. At the beginning of the twentieth century the 'Moscow school' of Mathematics was attempting to use rigorous arguments to establish evidence for the doctrine of free will, which *loosely* implies that each person is solely responsible for his or her actions [117]. During an era in which various political and economic ideologies were on the rise, the prospect of a widely held belief in such a doctrine posed serious threats to social order. P. A. Nekrasov, a mathematician from the Moscow school, used the quote at the beginning of this thesis as an argument for free will. He noted that averages of many people's behaviour (such as voting polls) apppeared to approach constant values, and claimed that this was evidence that decisions were being made independently, rather than under the influence of others [118]. Andrey Andreyevich Markov, an opponent to the Moscow school, developed the Markov chain to refute this claim, and indeed showed that independence was *not* a necessary condition for a Law of Large Numbers [74]. It is thrilling to consider that at this time abstract Probability formed the core of philosophical and political debates that came to have such a profound influence on the modern world.

3.2.1 Doeblin–Kolmogorov theory for countable state spaces

Although introduced by Markov in [74], the theory of chains on countable state spaces was mainly developed independently by Kolmogorov [61] and Doeblin [33] in the 1930s. Both were interested in classifying the long-time behaviour of chains. An important precursor to this section is the concept of an invariant distribution, meaning a probability measure $\pi(\cdot)$ for which

$$\pi(\cdot) = \pi P(\cdot). \tag{3.9}$$

Intuitively, if $X_i \sim \pi(\cdot)$, then $X_{i+m} \sim \pi(\cdot)$ for all $m \geq 0$. In this subsection we characterise the conditions under which $\pi(\cdot)$ both exists and is unique when $\mathbf{X}$ is countable, and establish when $\pi(\cdot)$ will be the *limiting* distribution, as in (3.7). To do this we must first introduce some *stability* concepts.

In the countable case, we say any two elements x and y *communicate*, denoted $x \leftrightarrow y$, if there are $n = n(x,y)$ and $m = m(y,x)$ such that $P^n(x,y) > 0$ and $P^m(y,x) > 0$. We can also express this idea through the notion of a *hitting time*

$$\tau_y = \inf\{t \geq 1 : X_t = y\},$$

calling state y 'reachable' from state x if either $\mathbb{P}_x[\tau_y < \infty] > 0$ or $y = x$. We define the communicating class of a state $x \in \mathbf{X}$ as $C(x) = \{y \in \mathbf{X} : y \leftrightarrow x\}$, which represents all the states that can eventually be reached from x. In fact, we can partition $\mathbf{X}$ into communicating classes, as "$\leftrightarrow$" defines an equivalence relation on $\mathbf{X}$ (see Appendix A), meaning we can write $\mathbf{X} = \bigcup_{i \in I} C(x_i)$ for some index set I, with $C(x_i) \cap C(x_j) = \emptyset$ for any $i \neq j$.

A Markov chain is called *irreducible* if $C(x) \equiv \mathbf{X}$, or equivalently $\mathbb{P}_x[\tau_y < \infty] > 0$ for all $x, y \in \mathbf{X}$. In words, any state can be reached from any other. If a chain is not irreducible any limiting distribution may critically depend on the starting point X_0.

Example 3.3. *Consider a chain with* $\mathbf{X} = \{1,2,3,4\}$*, and transition kernel defined by* $P(1,2) = P(2,1) = \theta_1$*,* $P(1,1) = P(2,2) = 1 - \theta_1$*,* $P(3,4) = P(4,3) = \theta_2$ *and* $P(3,3) = P(4,4) = 1 - \theta_2$*, for* $0 < \theta_i < 1$*. It can be shown that if* $x_0 = 1$ *or* 2 *then the limiting distribution is* $\pi = (1/2, 1/2, 0, 0)$*, while if* $x_0 = 3$ *or* 4 *then* $\pi = (0, 0, 1/2, 1/2)$*.*

Irreducibility implies any state *can* be reached from any starting point, but we would like to identify states that *will* be reached. In fact, in the countable case irreducible chains can be partitioned into two categories, which we call *recurrent* and *transient*. Since $\mathbb{P}_x[\tau_x < \infty] > 0$ for all $x \in \mathbf{X}$, these concepts are defined as

$$\mathbb{P}_x[\tau_x < \infty] = 1 \Rightarrow x \text{ is recurrent},$$
$$\mathbb{P}_x[\tau_x < \infty] < 1 \Rightarrow x \text{ is transient}.$$

Although this is a state-level definition, remarkably the categorisation is the same for all states in

the equivalence class $C(x)$ of x (shown in Appendix A). So in the irreducible case, if a single state is recurrent then the whole chain is, and similarly for transience.

We can also define recurrence of a state x through its *occupation time* as

$$\eta_x = \sum_{t=0}^{\infty} \mathbb{1}_x(X_t).$$

Recurrent states are those for which $\mathbb{E}_x[\eta_x] = \infty$. We say x is visited 'infinitely often.'

Proposition 3.4. *When* $\mathbf{X}$ *is countable,* $\mathbb{E}_x[\eta_x] = \infty \Leftrightarrow \mathbb{P}_x[\tau_x < \infty] = 1$.

Proof: See Appendix A.

In the case where $|\mathbf{X}| < \infty$, then irreducibility and transience cannot actually coincide.

Proposition 3.5. *If* $\{X_t\}_{t\geq 0}$ *unfolds on a finite state space, then irreducibility* $\Rightarrow$ *recurrence.*

Proof: A transient state $x \in \mathbf{X}$ will only be visited a finite number of times, so there exists some time T_x after which x will never be visited again. Where $C(x) = \mathbf{X}$, a single transient state $\Rightarrow$ every state is transient, so there would exist a time $T = \max\{T_x : x \in \mathbf{X}\}$ after which no state will be visited again. As the chain must be somewhere at this time, we have a contradiction, implying the chain must be recurrent. ■

When $|\mathbf{X}| = \infty$ an irreducible chain can be transient. A simple example is the random walk on $\mathbb{Z}$, with $X_0 = 0$ and transition kernel $P(x, x+1) = p$, $P(x, x-1) = 1-p$ for some $p \in [0,1]$. If $p = 1/2$ then it can be shown that $\mathbb{E}_0[\eta_0] = \infty$, but any other choice for p will result in a finite expected occupation time (for a proof see [86]). Since there are an infinite number of states, we cannot use the same argument to show recurrence as in the finite case.

We define the *period* of a state $x \in \mathbf{X}$ as

$$p_x = \gcd\{i \geq 1 : P^i(x,x) > 0\},$$

where gcd$A$ denotes the greatest common divisor of the set $A \in \mathbb{N}$. If $p_x = 1$ then x is called *aperiodic*. Periodicity is again a class property [86], so the states of an irreducible chain will all have the same period. A chain must be aperiodic for $P^m(x,\cdot)$ to converge to a limit. However, if we introduce

the 'average' kernel

$$A_m(x,\cdot) = \frac{1}{m}\sum_{i=1}^{m} P^i(x,\cdot),$$

then we can remove this requirement. The Markov chains considered here will be aperiodic, so we will only briefly mention A_m in what follows, though see [105] for more here.

Example 3.6. *Consider a chain with* $\mathbf{X} = \{1,2\}$*, and transition kernel defined by* $P(1,2) = P(2,1) = 1$ *and* $P(1,1) = P(2,2) = 0$*. If* $x_0 = 1$*, then* $P^{2m}(x_0,\cdot) = \delta_1(\cdot)$ *and* $P^{2m+1}(x_0,\cdot) = \delta_2(\cdot)$*, so* $P^m(x,\cdot)$ *never converges to a limit. However,* $A_m(x,\cdot)$ *will converge to* $(1/2,1/2)$*.*

The key results of this subsection are stated below.

Theorem 3.7. *If a Markov chain* $\{X_t\}_{t\geq 0}$ *is recurrent, then there exists a unique (up to a multiplicative constant)* σ*-finite*[4] *invariant measure.*

Proof: This is the first part of Theorem 10.0.1 of [81].

In the countable case this need not be a distribution (for example it could be Lebesgue measure). If the state space is finite, however, it will be, as the invariant measure will be finite and hence normalisable.

Theorem 3.8. *An irreducible, aperiodic, finite Markov chain has a unique invariant probability distribution.*

Proof: First note that if P is irreducible and aperiodic with real entries then the Perron–Frobenius theorem states that it has a unique largest real eigenvalue λ_1 with a unique left eigenvector, which can be chosen to have positive entries [91, 40]. As P is a square matrix, its left and right eigenvalues are the same [129]. It is clear that P has one as a right eigenvector, as

$$(P\mathbb{1})_i = \sum_j P(i,j) = 1,$$

where $\mathbb{1}$ denotes the column vector with each entry equal to one. So it is true that P also has a left eigenvector with eigenvalue one, which we can call π, meaning $\pi P = \pi$.

[4] A measure $\mu(\cdot)$ is called σ-finite if $\mathbf{X}$ can be written as a countable union $\bigcup_{i=1}^{\infty} A_i$, with $\mu(A_i) < \infty$ for each i.

To see that this is the largest eigenvalue, implying that π has positive entries and is unique (up to a multiplicative constant), note that any eigenvalue must satisfy $|\lambda_i| \leq 1$. To see this, assume x is a right eigenvector with corresponding eigenvalue λ_x. Take x_i as the largest element in the column vector x, and note that $Px = \lambda_x x$, meaning

$$\sum_i P(i,j)x_j = \lambda_x x_i.$$

The left-hand side is simply a weighted average of the elements of x, of which x_i is the largest, so taking absolute values gives

$$|\lambda_x||x_i| \leq |x_i| \implies |\lambda_x| \leq 1,$$

which completes the proof. ■

Kac's theorem gives a further characterisation, namely that $\pi(\{x\}) = \mathbb{E}_x[\tau_x]^{-1}$, relating the invariant distribution to the expected return time to a state $x \in \mathbf{X}$ [86]. This also sheds light on the countably infinite case, where we can introduce a further dichotomy. We say a recurrent chain is *positive* recurrent if the invariant measure is finite, and *null* recurrent if it is only σ-finite. Equivalently we can say that positive recurrent chains have finite expected return times, whereas for null recurrent chains these are infinite. More generally we call any Markov chain *positive* if it has an invariant probability measure. The last result in this subsection tells us that provided we can establish positivity, then we have all that is needed.

Theorem 3.9. *If an irreducible countable state Markov chain is positive, then it is recurrent. If it is aperiodic, then $\pi(\cdot)$ is also the limiting distribution for the chain.*

Proof: See Section 21.3, particularly Theorem 21.14, of [66].

3.2.2 Doeblin–Harris theory for general state spaces

Markov chains on general state spaces were first explored in detail by the pioneering work of Doeblin [34]. Later Harris [47] contributed significantly to the theory, leading some to refer to the 'Harris recurrence' school of Markov chains [28]. The works of Nummelin [88] and Meyn & Tweedie [81] (as well as others) have done much to refine and popularise the field.

Much of the theory from countable chains carries forward into the general case, but some modifications are required. The first of which is that if $\mathbf{X}$ is uncountably large, then irreducibility becomes

too strict a concept, since $P^n(x,y) = 0$ for any individual state $y \in \mathbf{X}$ by necessity. Without this we cannot define the equivalence relation "$\leftrightarrow$", so it becomes more difficult to decompose $\mathbf{X}$ as in the countable case (though see [127] for interesting discussion on this point). Fortunately, for our purposes little is lost by focusing instead on an analogue to irreducibility.

A chain $\{X_t\}_{t\geq 0}$ is called φ-*irreducible* with respect to some σ-finite measure $\varphi(\cdot)$ if for any $A \in \mathscr{B}$ with $\varphi(A) > 0$ we have:

$$\mathbb{P}_x[\tau_A < \infty] > 0, \tag{3.10}$$

for all $x \in \mathbf{X}$.

This is in some ways a more general condition, as irreducibility requires $\varphi(\{x\}) > 0$ for all $x \in \mathbf{X}$. In the countable case we can take $\varphi(\cdot)$ to be the counting measure $c(A) := |A|$, but in the general case this will not be σ-finite. One apparent drawback is that the choice of $\varphi(\cdot)$ seems arbitrary. It is comforting, therefore, to know that if a chain is φ-irreducible for some $\varphi(\cdot)$, then a *maximal* irreducibility measure $\psi(\cdot)$ exists, with the properties

- the chain $\{X_t\}_{t\geq 0}$ is ψ-irreducible
- if $\{X_t\}_{t\geq 0}$ is φ-irreducible, then $\varphi \ll \psi$, meaning $\psi(A) = 0 \Rightarrow \varphi(A) = 0$ for any $A \in \mathscr{B}$.

Note that what is important here is not so much the value of $\psi(A)$, but the collection of sets

$$\mathscr{B}^+ = \{A \in \mathscr{B} : \psi(A) > 0\},$$

which are the elements of $\mathscr{B}$ that can be reached from any choice of $\mu(\cdot)$. So there is actually an infinitely large family of equivalent maximal irreducibility measures $\psi(\cdot)$. Given an initial measure $\varphi(\cdot)$, we can actually construct one such $\psi(\cdot)$ using the transition kernel (see Appendix A).

Example 3.10. *Consider an $\{X_t\}_{t\geq 0}$ with transition kernel $P(1,1) = P(2,2) = \theta_1$ and $P(1,2) = P(2,1) = 1-\theta_1$. Then $\{X_t\}_{t\geq 0}$ is both $\delta_1(\cdot)$- and $\delta_2(\cdot)$-irreducible, and a maximal irreducibility measure is the counting measure $c(\cdot)$.*

Example 3.11. *Consider an $\{X_t\}_{t\geq 0}$ with $\mathbf{X} = \mathbb{R}$ and $\mathscr{B}$ the Borel σ-algebra on $\mathbb{R}$, with transition kernel $P(x,\cdot)$ a Gaussian distribution with mean x and variance 1, for all $x \in \mathbf{X}$. Then $P(x,A) > 0$ for any $A \in \mathscr{B}$ with $\mu^L(A) > 0$, for all $x \in \mathbf{X}$, where $\mu^L(\cdot)$ denotes Lebesgue measure on $\mathbb{R}$. So the chain is μ^L-irreducible.*

Recurrence in the general case is again more subtle, as the two definitions of a recurrent state which are equivalent in the countable case are no longer so here.

Proposition 3.12. *In the general case* $\mathbb{E}_x[\eta_A] = \infty \not\Rightarrow \mathbb{P}_x[\tau_A < \infty] = 1$, *for any* $A \in \mathscr{B}$ *and* $x \in A$.

Proof: A counterexample is given later in this section.

In general a set $A \in \mathscr{B}$ is called *recurrent* if for any $x \in A$ we have

$$\mathbb{E}_x[\eta_A] = \infty. \tag{3.11}$$

A set is called *uniformly transient* if there is a constant M which upper bounds this quantity. In the φ-irreducible case the entire chain is called recurrent if (3.11) holds for any $x \in \mathbf{X}$ and any $A \in \mathscr{B}^+$. It is transient if $\mathbf{X}$ can be covered by a countable collection of uniformly transient sets. A φ-irreducible chain will either be recurrent or transient (See Chapter 6 of [99]).

In the general case, Proposition 3.12 presents a problem. We resolve it by defining a stronger notion, known as *Harris* recurrence. A set $A \in \mathscr{B}$ is called Harris recurrent if

$$\mathbb{P}_x[\tau_A < \infty] = 1, \tag{3.12}$$

for all $x \in A$. The chain is called Harris recurrent (or simply Harris) if (3.12) holds for any $x \in \mathbf{X}$ and any $A \in \mathscr{B}^+$. The next example shows how the two definitions can result in practical problems.

Example 3.13. *Take* $P(x,A)$ *as a Harris recurrent kernel on* $\mathbf{X}$. *Now create a new chain on the extended space* $\mathbf{X}' := \mathbf{X} \cup N$, *where* $N = \{x_i\}_{i=1}^{\infty}$, *by setting* $P'(x,A) = P(x,A)$ *for all* $x \in \mathbf{X}$, $A \in \mathscr{B}^+$, *and* $P'(x_i, x_{i+1}) = q_i$, $P'(x_i, y) = 1 - q_i$ *for some specific* $y \in \mathbf{X}$. *So once the chain reaches* $\mathbf{X}$ *it remains there, and from each* x_i *the chain either moves to* $\mathbf{X}$ *or jumps to* x_{i+1}.

Under the dynamics of P, we have $\mathbb{P}_x[\tau_A < \infty] = 1$ *for any* $A \in \mathscr{B}^+$. *But provided* $\{q_i\}_{i=1}^{\infty}$ *is chosen such that* $1 \geq \prod_i q_i > 0$, *we have*

$$\mathbb{P}_{x_i}[\tau_y < \infty] = 1 - \prod_{j=i}^{\infty} q_j < 1.$$

Using this, we see that

$$\mathbb{P}_{x_i}[\tau_A < \infty] = \mathbb{P}_{x_i}[\tau_y < \infty]\mathbb{P}_y[\tau_A < \infty] = \mathbb{P}_{x_i}[\tau_y < \infty] < 1,$$

for any $A \in \mathscr{B}^+$. So the chain is not Harris recurrent. However, it is *recurrent (see Section 9.1.2 of [81]).*

The key point here in general is that if the chain is recurrent we can sometimes find a set $N \subset \mathbf{X}$ of potential starting points for $\{X_t\}_{t\geq 0}$ (with $\varphi(N) = 0$) from which the sets in $\mathscr{B}^+$ may not be visited. For a Harris chain this 'measure-theoretic pathology' (as it is called in [21]) is removed, so we can initialise the chain from any $x \in \mathbf{X}$.

Thankfully, the dichotomy of *positive* and *null* recurrent (and Harris recurrent) chains carries over from the countable case with no additional difficulties. In the general case a φ-irreducible chain is called *periodic* with period $p \geq 2$ if we can find a sequence of disjoint sets $\{S_1, ..., S_{p-1}\}$ such that for any $x \in S_i$

$$P(x, S_j) = 1 \text{ for } j = (i+1) \text{mod}(p).$$

Otherwise the chain is aperiodic [124]. The following two results conclude this subsection.

Theorem 3.14. *A φ-irreducible, aperiodic Markov chain has a unique σ-finite invariant measure.*

Proof: See Theorem 10.0.1 in [81].

Theorem 3.15. *If a φ-irreducible Markov chain $\{X_t\}_{t\geq 0}$ is positive, then it is recurrent.*

Proof: This is Proposition 10.1.1 of [81].

These results suit our purposes well. Specifically, if we can find an invariant probability distribution and can establish φ-irreducibility, then we know our chain is recurrent (though not necessarily Harris). Positive, φ-irreducible chains will therefore be our objects of study, since for these chains an invariant distribution both exists and is unique. However, we have not yet established whether the chain has a *limiting* distribution. We turn to this now.

3.2.3 Limiting distributions and ergodicity

The ultimate goal is to establish conditions under which we can approximate expectations by averaging across a Markov chain. Clearly (3.7) is a desirable property connected with this goal. In

this subsection we establish conditions for (3.7), and then in the next we connect these with (3.8) explicitly.

There are two different ways in which a Markov chain can be considered to converge in some sense to a limit. We can either consider the kernel itself $P^m(x,\cdot)$, or the average kernel $A_m(x,\cdot)$. Typically the word *ergodic* is attributed to the latter. A dynamical system in general is called *ergodic* if the average time spent in some set $A \in \mathscr{B}$ (in this case the long-run average from the chain) is equal to the 'spatial average' (in this case $\pi(A)$). Although Markov proved such results for Markov chains [74], Birkhoff and Von Neumann are credited with establishing them for more general dynamical systems, and beginning the field of Ergodic theory [12, 128]. Here, we choose to work with the convergence of P^m, which is a stronger notion, but provided chains are aperiodic nothing is lost by doing this [105].

We will define 'convergence' to $\pi(\cdot)$ using a distance metric on the space of probability measures over $\mathbf{X}$. Although several options exist [42], a choice which is both relatively well understood and strong enough for our purposes is *total variation*. For any two distributions $\mu(\cdot)$ and $\nu(\cdot)$ on $(\mathbf{X},\mathscr{B})$ this is defined as

$$\|\mu(\cdot)-\nu(\cdot)\|_{TV} := \sup_{A\in\mathscr{B}}\{\mu(A)-\nu(A)\}. \tag{3.13}$$

Intuitively, (3.13) gives the largest possible difference between the probability of any single event in $\mathscr{B}$ under $\mu(\cdot)$ and $\nu(\cdot)$. If both distributions admit densities, we can re-write (3.13) as

$$\|\mu(\cdot)-\nu(\cdot)\|_{TV} = \frac{1}{2}\int |\mu(x)-\nu(x)|dx, \tag{3.14}$$

which is proportional to the L_1 distance between $\mu(x)$ and $\nu(x)$ (we show the derivation explicitly in Appendix B). Removing the supremum can often make the distance easier to compute. Our metric $\|\cdot\|_{TV} \in [0,1]$, with $\|\cdot\|_{TV} = 1$ implying the distributions have disjoint supports.

Total variation convergence of a sequence of distributions $\{\varphi_n(\cdot)\}_{n\geq 0}$ to some limit $\varphi(\cdot)$ is written

$$\|\varphi_n(\cdot)-\varphi(\cdot)\|_{TV} \to 0$$

as $n\to\infty$. Two other common forms of convergence of probability measures are strong and weak, the former given by

$$\varphi_n(A)\to\varphi(A),\ \forall A\in\mathscr{B},$$

and the latter

$$\mathbb{E}_{\varphi_n}[f(X)]\to\mathbb{E}_{\varphi}[f(X)],$$

for all bounded continuous functions $f : \mathbf{X} \to \mathbb{R}$. Weak convergence of $\varphi_n(\cdot)$ is equivalent to convergence in distribution of the sequence of random variables $X_n \sim \varphi_n(\cdot)$. Strong $\Rightarrow$ weak, but not the reverse (a simple counter-example is the sequence $\delta_{\frac{1}{n}}(\cdot)$, which converges weakly to $\delta_0(\cdot)$ but not strongly). Total variation is in fact a stricter notion still as it implies a uniformity of convergence across sets in $\mathscr{B}$, whereas strong convergence only gives a pointwise result. With this machinery we can make our intuition concrete.

Definition. A Markov chain $\{X_t\}_{t\geq 0}$ with transition kernel P and invariant distribution $\pi(\cdot)$ is called *ergodic* if

$$\|\mu P^n(\cdot) - \pi(\cdot)\|_{TV} \to 0$$

as $n \to \infty$, for any initial distribution $\mu(\cdot)$.

Definition. A Markov chain $\{X_t\}_{t\geq 0}$ with transition kernel P and invariant distribution $\pi(\cdot)$ is called *a.s. ergodic* if

$$\|\mu P^n(\cdot) - \pi(\cdot)\|_{TV} \to 0$$

as $n \to \infty$, from π-*almost-any* starting point.

The phrase π-*almost-any* means that the set S of starting points for which the chain is not ergodic is such that $\pi(S) = 0$. Given the stability structures introduced in the previous sections, we can now identify ergodic Markov chains on general state spaces.

Theorem 3.16. (Aperiodic ergodic theorem). *A φ-irreducible, aperiodic, positive Harris recurrent Markov chain is ergodic.*

Proof: See Theorem 3.1 of [124].

Although we are mainly interested in the general case, the following are interesting to note.

Corollary 3.17. *A countable state Markov chain which is irreducible, aperiodic and positive is ergodic.*

Corollary 3.18. *A finite state Markov chain which is irreducible and aperiodic is ergodic.*

The first follows since positivity $\Rightarrow$ recurrence, and recurrence $\Rightarrow$ Harris in the countable case. In the finite case irreducible $\Rightarrow$ positive in addition.

In practice, the most difficult to establish of the conditions required for the above theorem is Harris recurrence. Positivity is relatively easy, we just find a $\pi(\cdot)$ that satisfies (3.9). Likewise it is usually clear from P whether the chain will be φ-irreducible and aperiodic. Conditions for Harris recurrence are outlined in [124]. However, it is useful to note that in the absence of Harris recurrence we can still rely on a slightly weaker result.

Theorem 3.19. *If $\{X_t\}_{t\geq 0}$ is a φ-irreducible, aperiodic, positive recurrent Markov chain with initial distribution $\delta_{x_0}(\cdot)$, so that $\mu P^t(\cdot) = P^t(x_0,\cdot)$ then $\{X_t\}_{t\geq 0}$ is a.s. ergodic.*

Proof: See e.g. Theorem 4 of [105].

In short, provided we know the support for $\pi(\cdot)$, we can choose a starting point for the chain within this support, and we no longer need Harris recurrence to ensure convergence.

3.2.4 Limit theorems & geometric ergodicity

If we can establish that a Markov chain is ergodic or a.s. ergodic, then we have the Law of Large Numbers result that we desire.

Theorem 3.20. (Law of Large Numbers for Markov chains). *If a chain $\{X_t\}_{t\geq 0}$ is ergodic, then for any $f : \mathbf{X} \to \mathbb{R}$ with $\mathbb{E}_\pi[f(X)] < \infty$, the estimator*

$$\tilde{f}_m \to \mathbb{E}_\pi[f(X)]$$

with probability one, as $m \to \infty$.

Proof: See the corollary to Theorem 3.6 in Chapter 4 of [97].

In the case of a.s. ergodicity, we require the chain to be initiated from within the support of $\pi(\cdot)$.

Of course, consistency is a desirable property of an estimator. However, we would really like to say more than this about $\tilde{f}_m$. To do so, however, we need stricter conditions on $\{X_t\}_{t\geq 0}$. A condition which is sufficient (though not always necessary) is called *geometric ergodicity*. We first motivate

why this is such a desirable property for the chain, before formally defining it and showing how it can be established in the next section.

Theorem 3.21. (Markov Chain Central Limit Theorem). *If a Markov chain $\{X_t\}_{t\geq 0}$ with transition kernel P is geometrically ergodic, and $f : \mathbf{X} \to \mathbb{R}$ is a Borel functional with $\mathbb{E}_\pi[|f|^{2+\varepsilon}] < \infty$ for some $\varepsilon > 0$, then*

$$\sqrt{m}\left(\tilde{f}_m - \mathbb{E}_\pi[f(X)]\right) \xrightarrow{d} N\left(0, v(P,f)\right), \tag{3.15}$$

as $m \to \infty$, where $v(P,f)$ depends on both the functional f and the transition kernel P. If the chain is reversible *(defined in Chapter 4), then (3.15) holds provided $\mathbb{E}_\pi[|f|^2] < \infty$.*

Although still an asymptotic property, clearly this result gives us considerably more confidence in our estimator $\tilde{f}_m$, particularly if we can get a reasonable idea of $v(P,f)$.

The form of $v(P,f)$ is in fact quite intuitive. To motivate the derivation, consider a sequence of m random variables $\{X_1, ..., X_m\}$, each with marginal distribution $\pi(\cdot)$, with $\text{Var}[X_i] < \infty$, and the estimator

$$\hat{f}_m = \frac{1}{m}\sum_{i=1}^{m} f(X_i).$$

This could be the first n random variables in a stationary Markov chain, but we do not have to assume this dependence structure. From basic properties of expectation and variance we can conclude that $\mathbb{E}[\hat{f}_m] = \mathbb{E}_\pi[f(X)]$, and

$$\text{Var}[\hat{f}_m] = \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m} \text{Cov}_\pi[f(X_i), f(X_j)].$$

If we make the additional assumption that $\text{Cov}[f(X_i), f(X_{i+k})]$ is independent of i (which is true for a stationary Markov chain), then we can re-write this as

$$\text{Var}[\hat{f}_m] = \frac{1}{m}\left(\text{Var}_\pi[f(X_1)] + 2\sum_{k=1}^{m-1}\left(\frac{m-k}{m}\right)\text{Cov}_\pi[f(X_1), f(X_{1+k})]\right). \tag{3.16}$$

Noting that for any fixed k, the ratio $(m-k)/m \to 1$ as $m \to \infty$, we can see that

$$\lim_{m\to\infty} m\text{Var}[\hat{f}_m] = \text{Var}_\pi[f(X_1)] + 2\sum_{k=1}^{\infty}\text{Cov}_\pi[f(X_1), f(X_{1+k})]. \tag{3.17}$$

The remarkable result for Markov chains is that (3.17) can still be used even if the sequence is not stationary, provided the chain is geometrically ergodic.

Equation (3.17) can also be used to 'choose between' different Markov chains as a means for constructing estimators. Clearly the objective is to minimise $\sum_{k=1}^{\infty} \mathrm{Cov}_{\pi}[f(X_1), f(X_{1+k})]$. Ranking different Markov chains in such a way is called a *Peskun ordering* [92, 125].

Now that we have motivated the concept, we turn to a full definition.

Definition. A Markov chain $\{X_t\}_{t\geq 0}$ with transition kernel P and invariant distribution $\pi(\cdot)$ is called *geometrically ergodic* if

$$\|\mu P^m(\cdot) - \pi(\cdot)\|_{TV} \leq M(\mu) r^m, \tag{3.18}$$

for some function $M \geq 0$ and some $0 \leq r < 1$, for any initial distribution $\mu(\cdot)$.

As (3.18) comes from a bound that decreases *geometrically* with m, the etymology is fairly unambiguous. Less clear is the case where M is bounded above, which is termed *uniform* ergodicity. Uniform ergodicity means that the distance between $P^m(x_0, \cdot)$ and $\pi(\cdot)$ is decreasing geometrically in m, and that a bound exists which is uniform for any choice of x_0.

The intuition for a geometric bound comes from the case where $\mathbf{X}$ is countable. We give an illustrative example in the finite case here, where P is an $s \times s$ matrix. If we assume P is symmetric, irreducible and aperiodic, then by the spectral theorem (e.g. [112]) we can write it in a diagonal form

$$P = U^T D U,$$

where each column of U is a *left* eigenvector of P, and $D = \mathrm{diag}(\lambda_1, ..., \lambda_s)$ is a diagonal matrix of real eigenvalues [126]. We have already shown that the largest eigenvalue is one. In fact it is also true that $|\lambda_i| < 1$ for all other eigenvalues (see Chapter 12 of [66]). The quantity

$$\lambda_G = 1 - \sup_{|\lambda_i|<1} |\lambda_i|,$$

is known as the *spectral gap* for the chain. It can be shown that the existence of a spectral gap $\lambda_G > 0$ is equivalent to the notion of geometric ergodicity [102]. If we write the eigenvalues in descending order, with $\lambda_1 = 1$, then it is actually true that any initial distribution vector μ can be written

$$\mu = \pi + \sum_{i=2}^{s} a_i e_i,$$

for some constants $a_i \in \mathbb{R}$, where e_i denotes the ith eigenvector of P (see Chapter 12 of [66]). If we apply P to the right-hand side (giving the distribution for X_1 as μP), then we have $\mu P =$

$\pi + \sum_{i=2}^{s} a_i \lambda_i e_i$. Iterating gives

$$\mu P^m = \pi + \sum_{i=2}^{s} a_i \lambda_i^m e_i.$$

Because $|\lambda_i| < 1$ for all $i \geq 2$, then as $m \to \infty$

$$\|\mu P^m - \pi\|_{TV} \leq \sum_{i=2}^{s} |a_i||\lambda_i|^m \|e_i\|_{TV} = O\left((1-\lambda_G)^m\right),$$

so convergence here occurs at a geometric rate. A natural question is when such a rate holds in the general case, and how to establish a bound such as (3.18) in practice. Although the same spectral decomposition could be applied to general state transition kernels (provided a suitable inner product is defined), it can often be extremely difficult to find the eigenvalues of P. We instead discuss a different approach here.

3.2.5 Establishing geometric ergodicity

A very useful result, again first introduced by Doeblin (see the Appendix of [69]), is the *coupling* inequality. A coupling of any two random variables $X \sim \mu(\cdot)$ and $Y \sim \nu(\cdot)$ is any joint distribution $\Lambda(\cdot)$ for (X,Y) such that the marginals for X and Y are $\mu(\cdot)$ and $\nu(\cdot)$. A simple example is the case where X and Y are both $N(0,1)$, in which case any bivariate Gaussian distribution $N(0,\Sigma)$ with $\Sigma_{11} = \Sigma_{22} = 1$ defines a coupling of X and Y. The choice of how X and Y depend on each other is free provided the marginals are preserved. Another coupling in this case would be $X = Y$ with probability one.

Theorem 3.22. (Coupling Inequality) *For any coupling $\Lambda(\cdot)$ of random variables $X \sim \mu(\cdot)$ and $Y \sim \nu(\cdot)$ we have*

$$\|\mu(\cdot) - \nu(\cdot)\|_{TV} \leq \mathbb{P}_{\Lambda}[X \neq Y]. \tag{3.19}$$

Proof: See Appendix A.

Example 3.23. *Consider the case $\mu(\cdot) = \nu(\cdot)$, so that $\|\mu(\cdot) - \nu(\cdot)\|_{TV} = 0$.*

- *One coupling $\Lambda_1(\cdot)$ could be defined such that X and Y are independent, meaning $P_{\Lambda_1}[X \neq Y] = 1$, giving a very loose bound*
- *Another, $\Lambda_2(\cdot)$, could be that $X = Y$ with probability 1, so that we sample $X \sim \mu(\cdot)$ and then set Y to be the same value. In this case $P_{\Lambda_2}[X \neq Y] = 0$, so that the bound on $\|\mu(\cdot) - \nu(\cdot)\|_{TV}$ is saturated.*

Note that in both cases the marginal distributions for both X and Y would be $\mu(\cdot)$, or equivalently $\nu(\cdot)$.

To return now to the goal, we seek a bound on the distance $\|P^m(x_0,\cdot)-\pi(\cdot)\|_{TV}$ which decreases geometrically in m. With the coupling inequality at our disposal, we can construct such a bound if we can find a coupling $\Lambda(\cdot)$ of $X_m \sim P^m(x_0,\cdot)$ and $Y \sim \pi(\cdot)$ such that $\mathbb{P}_\Lambda[X_m \neq Y] \propto r^m$ for some $r<1$. To construct such a $\Lambda(\cdot)$, we must introduce some further concepts.

We say that a set $C \in \mathbf{X}$ is *small* if there exists $m_0 \in \mathbb{N}$ and $\varepsilon > 0$ such that for any $A \in \mathscr{B}$

$$P^{m_0}(x_0,A) \geq \varepsilon\psi(A),\ \forall x_0 \in C, \tag{3.20}$$

where $\psi(\cdot)$ is some probability measure. Equation (3.20) is called a *minorisation* condition (see Section 5.2 of [81]). In the countable case it is also known as *Doeblin's condition* [86].

We focus on the case $m_0 = 1$ here, for ease of exposition, though the extension to any finite m is straightforward [105]. If (3.20) holds for some set C, then whenever $\{X_t\}_{t\geq 0}$ is in C we can 'split' the transition kernel P into two constituent parts, one of which does not depend on the current position in the chain, using the decomposition

$$P(x_0,\cdot) = \varepsilon\psi(\cdot) + (1-\varepsilon)R(x_0,\cdot),$$

where $R(x_0,\cdot) = (P(x_0,\cdot)-\varepsilon\psi(\cdot))/(1-\varepsilon)$ is also a transition kernel. Generating the next sample in a Markov chain with kernel P can therefore be done in two stages, whenever the current point in the chain is some $x \in C$. First, draw a Bernoulli random variable with probability of success ε, and then conditional on success draw from $\psi(\cdot)$, otherwise draw from $R(x,\cdot)$. Note that the *marginal* transition kernel is still P. The random times T at which draws are made from $\psi(\cdot)$ are known as *regeneration times* for the chain. At these points the next sample is drawn completely independently of even the current value in the chain. This 'splitting' construction was introduced independently by Nummelin [87] and Athreya & Ney [4].

Example 3.24. *Consider a Markov chain with state space $\mathbf{X} = [0,10]$ and where the transition kernel $P(x,\cdot)$ is a standard Gaussian distribution centred at x and truncated at 0 and 10, with density $p(y|x)$. Then for any $A \in \mathscr{B}$ we have*

$$P(x,A) = \int_A p(y|x)dx \geq c_{min}\int_A dy,$$

where

$$c_{min} = \inf_{x,y} p(y|x).$$

Hence we can set $\varepsilon = 10c_{min}$*, giving*

$$P(x,A) \geq \varepsilon \tilde{\mu}^L(A),$$

where $\tilde{\mu}^L(A) = \mu^L(A \cap \mathbf{X})/10$ *is the uniform distribution over* $\mathbf{X}$*. So here the whole state space is a small set.*

This concept of regeneration allows us to find a geometric bound using (3.19). Specifically, we consider two Markov chains $\{X_t\}_{t\geq 0}$ and $\{Y_t\}_{t\geq 0}$, and consider a sequence of couplings on the pairs (X_m, Y_m), such that $\mathbb{P}[X_m \neq Y_m]$ decreases geometrically in m. If $\{Y_t\}_{t\geq 0}$ is initialised at stationarity (i.e. $Y_0 \sim \pi(\cdot)$), then we have the bound we seek. The concrete construction in the case $m_0 = 1$ is:

1. Initialise two Markov chains, both with transition kernel P and invariant distribution $\pi(\cdot)$. For the chain $\{X_t\}_{t\geq 0}$ we set $X_0 = x_0$, and for $\{Y_t\}_{t\geq 0}$ we set $Y_0 \sim \pi(\cdot)$

2. At each iteration m, if $(x_{m-1}, y_{m-1}) \notin C \times C$ then we draw $X_m \sim P(x_{m-1}, \cdot)$ and $Y_m \sim P(y_{m-1}, \cdot)$ independently

3. But, in the case $(X_{m-1}, Y_{m-1}) \in C \times C$, we draw $U_m \sim \text{Bernoulli}(\varepsilon)$. If $U_m = 0$ then we draw $X_m \sim R(x_{m-1}, \cdot)$ and $Y_m \sim R(y_{m-1}, \cdot)$ independently, but if $U_m = 1$ we set $X_m = Y_m \sim \psi(\cdot)$, and draw subsequent values for each chain such that they remain equal

To see how such a construction induces a geometric bound on $\|P^m(x_0, \cdot) - \pi(\cdot)\|_{TV}$, it is important to note that since $Y_0 \sim \pi(\cdot)$, each Y_m has marginal distribution $\pi(\cdot)$. First consider the case $C = \mathbf{X}$, so that $(x_{m-1}, y_{m-1}) \in C \times C$ at every iteration. In this instance

$$\|P^m(x_0, \cdot) - \pi(\cdot)\|_{TV} \leq \mathbb{P}[X_m \neq Y_m] = (1-\varepsilon)^m.$$

In fact, this bound is uniform for any $x_0 \in \mathbf{X}$, so the case $C \in \mathbf{X}$ corresponds to a uniformly ergodic chain.

Theorem 3.25. *An irreducible, aperiodic finite state Markov chain is uniformly ergodic.*

Proof: Irreducibility implies that for any $(x,y) \in \mathbf{X} \times \mathbf{X}$ there is an $n_0 = n_0(x,y)$ such that $P^{n_0}(x,y) > 0$. We state a fact from number theory that since there are a finite number of (x,y) pairs then there is an n such that $P^n(x,y) > 0$ for all of them (see Lemma 1.27 on page 20 of [66] for a proof). Take

$$\delta = \inf\{P^n(x,y) : x,y \in \mathbf{X}\},$$

and define $c^*(\cdot) = c(\cdot)/c(\mathbf{X})$, where $c(A) = |A|$. Then $c^*(\cdot)$ is a probability measure over $\mathbf{X}$ and

$$P^n(x,A) \geq \delta c^*(A),$$

for any $A \in \mathscr{B}$ and any $x \in \mathbf{X}$. ■

Note that in the countable case this approach will no longer yield a proof, as $\inf\{P^n(x,y) : x,y \in \mathbf{X}\}$ can be zero. In fact, typically in the case of an unbounded $\mathbf{X}$, a φ-irreducible, aperiodic, positive Harris chain will not be uniformly ergodic. It is more common, therefore, to seek a small set C which is a *proper* subset of $\mathbf{X}$. In this case, the additional concern is how often $(x,y) \in C \times C$ in the split chain construction. To ensure that this happens often enough to construct a geometric bound on the coupling time, we need to consider how often the chains will visit C.

Example 3.26. *Consider a Markov chain with transition kernel $P(x,dy) = p(y|x)dy$ and state space $\mathbf{X} = \mathbb{R}$, which is ergodic to some distribution $\pi(\cdot)$. Note that unlike in Example 3.24, we now have*

$$\inf_{x,y \in \mathbf{X}} p(y|x) = 0,$$

so we can no longer lower bound $p(y|x)$ over all $x,y \in \mathbf{X}$ with some positive constant to create a minorisation condition. However, if we take the set $C = [0,10]$ and $\tilde{\mu}^L(\cdot)$ as the uniform distribution over $[0,10]$ (i.e. a distribution that only *has support in the set C), then we can still find an ε such that*

$$P(x,A) \geq \varepsilon \tilde{\mu}^L(A),$$

for any $x \in C$ and $A \in \mathscr{B}$. Here

$$\varepsilon = \inf_{x,y \in C} p(y|x),$$

which will be positive and finite provided $p(y|x)$ is bounded away from 0 and ∞ on $[0,10]$.

We only offer intuition for how to construct a geometric bound here, closely following that given in [57]. We need to consider the distribution of τ_C, the return time to C, defined as

$$\tau_C = \min\{m \geq 1 \ : \ X_m \in C \mid X_0 \in C\}.$$

If each chain spends enough time in C then we should have enough opportunities for $\{X_t\}_{t\geq 0}$ and $\{Y_t\}_{t\geq 0}$ to 'coalesce' (i.e. become equal) so that we can still establish a geometric bound. The essential (and intuitive) requirement is that for any $x \in C$, τ_C follows a distribution which has tails at least as light as a geometric random variable. Mathematically we need to show that $\mathbb{E}_x[e^{\beta_1 \tau_C}]$ exists for some $\beta > 0$, since

$$\mathbb{E}_x[e^{\beta \tau_C}] = \sum_{t=1}^{\infty} e^{\beta t} \mathbb{P}_x[\tau_C = t] < \infty$$

implies that the probability $\mathbb{P}_x[\tau_C = t] \in o(e^{-\beta t})$.

If it can be established that the return times to C have geometric tails, then we can still construct a bound on the total variation distance which decays geometrically in m [57, 81]. Establishing that τ_C has such tails is often most easily done through the use of a *drift* condition. We find some function $V : \mathbf{X} \to [1, \infty)$ for which $V(x) \to \infty$ as $|x| \to \infty$ (we call such a function *coercive*), and some $\lambda < 1$ and $b \leq \infty$ for which

$$PV(x) \leq \lambda V(x) + b\mathbb{1}_C(x), \ \ \forall x \in \mathbf{X}. \tag{3.21}$$

If such a *Lyapunov* function can be found, we can fix a small set to be $C = \{x \in \mathbf{X} : V(x) \leq d\}$. We can then construct a bound of the form

$$\|P^m(x_0, \cdot) - \pi(\cdot)\|_{TV} \leq MV(x_0)r^m,$$

for some $r < 1$ and $M < \infty$. Since V is unbounded above, the bound is not uniform, but it does satisfy the requirements of (3.18).

The reason we require $V \geq 1$ is that the original results demonstrating its use in establishing convergence bounds in fact are stronger than those we discuss here. In Chapter 16 of [81], it is shown that establishing both (3.21) and (3.20) provides a geometric bound on the so called V-norm distance between $P^m(x_0, \cdot)$ and $\pi(\cdot)$, given by

$$\|\mu(\cdot) - \nu(\cdot)\|_V := \sup_{f \in F} |\mathbb{E}_\mu[f(X)] - \mathbb{E}_\nu[f(Y)]|,$$

where $F = \{f : \mathbf{X} \to \mathbb{R} : |f(x)| \leq V(x), \ \forall x \in \mathbf{X}\}$. It is straightforward to see that total variation distance is the special case where $V(x) \equiv 1$ (see [105] for a proof). So for any $V \geq 1$, we have

$$\|\mu(\cdot) - \nu(\cdot)\|_{TV} \leq \|\mu(\cdot) - \nu(\cdot)\|_V$$

Because of this, geometric ergodicity is sometimes discussed along with V-uniform ergodicity, which is an equivalent bound on the V-norm distance [81, 102].

3.2.6 Central Limit Theorems from geometric ergodicity

A geometric bound is not always necessary to establish Central Limit Theorems (CLTs) for Markov chain estimators (in many case polynomial bounds suffice, see e.g. [39]). But proving the existence of (3.18) allows CLTs to be found in some generality, without *too* much difficulty. Usual proofs rely on the existence of solutions to *Poisson* equations, e.g. [105]. We instead give some intuition using an approach that more naturally connects with the classical result for independent and identically distributed (iid) random variables, taken primarily from [45, 27]. First we require some further definitions.

Definition. The *characteristic function* of a random variable[5] X is the function $\varphi_X : [0,\infty) \times \mathbf{X} \to \mathbb{C}$ given by

$$\varphi_X(t) = \mathbb{E}[e^{itX}]. \tag{3.22}$$

In fact, the characteristic function always exists, and completely characterises the distribution of X, in the sense that two random variables X and Y have the same distribution if and only if $\varphi_X(t) = \varphi_Y(t)$ for all t (see e.g. Corollary B.106 on page 645 of [114]).

We will use this alternative method of analysing random variables to prove Central Limit Theorems, first for independent and identically distributed random variables, and then in the Markovian case. A useful property of $\varphi_X(t)$ here is that for any $a, b \in \mathbb{R}$ and any independent random variables X and Y on $(\mathbf{X}, \mathscr{B})$ we can calculate the characteristic function of the linear combination $aX + bY$ as

$$\varphi_{aX+bY}(t) = \mathbb{E}[e^{it(aX+bY)}] = \mathbb{E}[e^{itaX}e^{itbY}] = \varphi_X(at)\varphi_Y(bt).$$

Denoting $\mu_X(\cdot)$ as the distribution for X, another identity is

$$\mathbb{E}[X^k] = (-i)^k \varphi_X^{(k)}(0),$$

as

$$\varphi_X^{(k)}(t) = \frac{d^k}{dt^k}\int e^{itx}\mu_X(dx) = \int \frac{\partial^k}{\partial t^k} e^{itx}\mu_X(dx) = i^k \int x^k e^{itx}\mu_X(dx).$$

The characteristic function is a more general form of the moment generating function $M(t) = \mathbb{E}[e^{tX}]$. Note that $\varphi_X(t)$ always exists whereas for $M(t)$ this depends on the tails of $\mu_X(\cdot)$.

[5] We restrict the definition here to one dimensional random variables, though extensions are straightforward.

We also recall the definition of convergence in distribution for a sequence of random variables $\{X_1, .X_2, ...\}$. We write

$$X_n \xrightarrow{d} X$$

if for the sequence of cumulative distribution functions given by $F_n(x) = \mathbb{P}[X_n \leq x]$ we have $F_n(x) \to F(x)$ as $n \to \infty$, at every x for which F is continuous, where $F(x) = \mathbb{P}[X \leq x]$ for some random variable X. The final component needed for our goal is the following theorem.

Theorem 3.27. (Lévy continuity theorem). *Suppose we have a sequence of random variables* $\{X_1, X_2, ...\}$*, with corresponding characteristic functions* $\{\varphi_{X_1}(t), \varphi_{X_2}(t), ...\}$*. If the sequence* $\{\varphi_{X_n}(t)\}_{n=1}^{\infty}$ *converges pointwise to a limit, i.e.*

$$\varphi_{X_n}(t) \to \varphi_X(t),$$

for all $t \in \mathbb{R}$*, then* $\varphi_X(t)$ *is the characteristic function of some random variable* X*, and*

$$X_n \xrightarrow{d} X.$$

Proof of this is given for example in Section B.4.2 on page 640 of [114].

Theorem 3.28. (Central Limit Theorem for iid sequences). *If we write*

$$S_n = \frac{1}{n} \sum_{i=1}^{n} f(X_i)$$

to denote the sample average of n independent and identically distributed random variables $f(X_1)$*,* $f(X_2)$*,...,*$f(X_n)$ *each with* $\mathbb{E}[f(X_i)] = \mu$ *and* $Var[f(X_i)] = \sigma^2 < \infty$*, then*

$$\sqrt{n}(S_n - \mu) \xrightarrow{d} N(0, \sigma^2).$$

Proof: Using properties of characteristic functions, we have

$$\varphi_{\sqrt{n}(S_n-\mu)}(t) = \varphi_{\frac{1}{\sqrt{n}}\sum_i (f(X_i)-\mu)}(t) = \left[\varphi_{f(X_i)-\mu}\left(\frac{t}{\sqrt{n}}\right)\right]^n$$

Also note that the Taylor series expansion of $\varphi_{f(X)}(t)$ about $t = 0$ is

$$\varphi_{f(X)}(t) = 1 + i\mathbb{E}[f(X)]t - \mathbb{E}[f(X)^2]t^2/2 + o(t^2), \;\; t < 1.$$

Here we have $\mathbb{E}[f(X_i)-\mu]=0$ and $\mathbb{E}[(f(X_i)-\mu)^2]=\sigma^2$. Combining the two expressions gives

$$\varphi_{\sqrt{n}(S_n-\mu)}(t)=\left[1-\frac{\sigma^2t^2}{2n}+o(t^2/n)\right]^n\to e^{-\sigma^2t^2/2},$$

as $n\to\infty$, which is the characteristic function of a $N(0,\sigma^2)$ random variable. ■

In the case where $\{X_t\}_{t\geq 0}$ is a Markov chain, then the characteristic function of $\sqrt{n}(S_n-\mu)$ does not have the same representation as a product, so such a straightforward proof is not possible. To deduce something similar, we recall the operator $Pf(x)=\int f(y)P(x,dy)$ and define a generalisation

Definition. The operator-valued generating function of a Markov chain $\{X_t\}_{t\geq 0}$ is defined as

$$P_{it}f(x)=\int e^{ity}f(y)P(x,dy),$$

for any measurable $f:\mathbf{X}\to\mathbb{C}$.

Clearly we have $P_0=P$. More generally, if we denote by $\psi_n(t)$ the characteristic function of $Y_n=f(X_1)+f(X_2)+\ldots+f(X_n)$, and $1(x)$ the function which maps any point $x\in\mathbf{X}$ to 1, then if the chain is started at $X_0=x$ we have

$$\psi_n(t)=\mathbb{E}[e^{itY_n}]=\mathbb{E}[e^{it\Sigma f(X_i)}]=\int\ldots\int e^{itf(x_1)}\ldots e^{itf(x_n)}P(x_0,dx_1)\ldots P(x_{n-1},dx_n)=P_{it}^n1(x).$$

We can see therefore that for Markov chains the operator-valued generating function does have a product representation.

As has already been alluded to, geometric ergodicity ergodicity of a Markov chain is shown in [102] to be equivalent to fact that the operator P has an L^2 spectral gap, meaning

$$\|P\|_{L^2(\pi)}:=\sup_{f\in\mathscr{F}}\int\left(f(y)P(x,dy)\right)^2\pi(dx)<1, \tag{3.23}$$

where $\mathscr{F}:=\{f:\mathbb{E}_\pi[f]=0,\mathbb{E}_\pi[f^2]<\infty\}$ (see [102] for details). If this property holds, then it is in fact possible to show that for t close to 0

$$\psi_n(t)=\lambda(it)^n(1+t\theta_1(t))+\rho_2^nt\theta_2(n,t),$$

where $\lambda(it)$ is the largest (in absolute value) eigenvalue of P_{it}, $\theta_1(t)$ and $\theta_2(n,t)$ are bounded and $0<\rho_2<1$. It is also the case when (3.23) holds that $\lambda(it)$ can be written as

$$\lambda(it)=1+it\mathbb{E}_\pi[f(X)]-v(P,f)t^2/2+O(t^3).$$

With this representation, similar arguments to those used in the classical Central Limit Theorem show that

$$\sqrt{n}(S_n - \mu) \xrightarrow{d} N(0, v(P, f)),$$

as required. The full argument is given in [45].

3.2.7 Geometric ergodicity on computers

As far as mathematical objects are concerned, we have now established that geometric ergodicity is a desirable property for general state space Markov chains, and that chains which are simply ergodic may not converge at a geometric rate. We have also, however, established in Subsection 3.2.5 that any *finite* ergodic Markov chain *will* be geometrically ergodic (in fact uniformly so over any starting point $x \in \mathbf{X}$). In Section 3.1 we have discussed the benefits and limitations of working with continuous random variables, and it is worth at this point having a similar discussion regarding general state space Markov chains.

Any Markov chain that is being simulated on a computer will necessarily be finite. The chains we simulate when performing Markov chain Monte Carlo are finite approximations to general state space chains, owing to the finite memory restrictions of computers. So any Markov chain Monte Carlo method, in practice, will converge to its equilibrium distribution at a geometric rate. With this in mind, it is important to understand what is being gained by establishing whether the general state space object that is being approximated will also be geometrically ergodic.

A short experiment illustrates why establishing geometric bounds 'in the limit' has some value on finite state spaces. Consider two Markov chains defined on the positive integers (a countably infinite state space):

1. A simple aperiodic random walk model, with $P(i, i-1) = P(i, i) = P(i, i+1) = 1/3$ for $i \geq 2$ and $P(1,1) = 2/3$, $P(1,2) = 1/3$

2. A slight variation in which $P(i, i-1) = 2/6$, $P(i, i) = 3/6$ and $P(i, i+1) = 1/6$ for $i \geq 2$ and $P(1,1) = 5/6$, $P(1,2) = 1/6$.

It is straightforward to see that both chains are irreducible and aperiodic. The first is in fact only null

recurrent, with the counting measure $c(\cdot)$ invariant, as for any j

$$\sum_{i=1}^{\infty} c(\{i\})P(i,j) = \sum_{i=1}^{\infty} P(i,j) = \sum_{i=1}^{\infty} P(j,i) = 1 = c(\{j\}).$$

The second is positive, with the geometric distribution as invariant measure. It has been proven that model 2 produces a geometrically ergodic chain in the countably infinite case [78], whereas the first is not even ergodic.

We consider finite state approximations to both models, with both approximations approaching the truth as the dimension $n \to \infty$. In each case we let n grow and compute the second largest eigenvalue (in absolute terms) from the transition matrix P. Figure 3.1 shows the results.

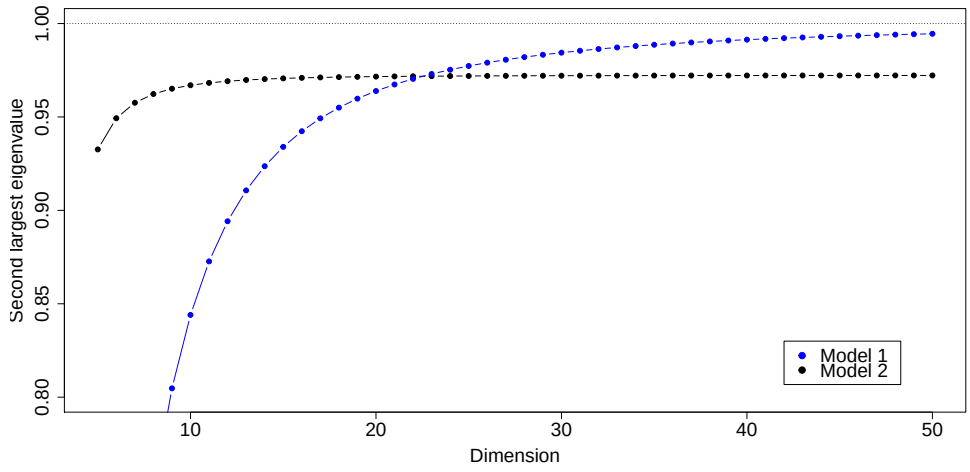

Figure 3.1: The second largest absolute eigenvalue plotted against state space dimension for two simple Markov chains.

If we denote the second largest eigenvalue from model i in dimension n by $\lambda_{i,n}$. it is clear numerically that $\lambda_{1,n} \to 1$ as $n \to \infty$, whereas $\lambda_{2,n} \to \lambda_2 < 1$. The point of the example is to highlight that geometric ergodicity in the limiting case of an infinite state space implies that the spectral gap, or equivalently the geometric rate of convergence, is robust to increasing dimension. So as in the continuous random variables case, if we develop analytical results for Markov chains in the general state space case, then the analysis will be robust to any increasing state space size n, which is clearly desirable.

3.2.8 Qualitative and quantitative bounds

Geometric ergodicity is often referred to as a *qualitative* bound. Since we usually do not know the constants M and r, all we know is that a bound exists which decreases geometrically as the chain evolves. Thankfully this is all that is required for Central Limit Theorems to exist for Markov chain estimators of interest to us.

There is still, however, considerable interest in developing non-asymptotic bounds, both for the distance $\|P^m(x_0,\cdot) - \pi(\cdot)\|_{TV}$, and some loss function associated with the estimator $\tilde{f}_m$. The former is of assistance when comparing different Markov chains, while the latter provides a more direct assessment of the quality of the estimator $\tilde{f}_m$ after some finite number of samples m. Clearly both are highly desirable, and predictably both are very challenging to show.

Quantitative bounds on the total variation distance between $P^m(x_0,\cdot)$ and $\pi(\cdot)$ exist (e.g. [57, 105]), based on the drift and minorisation techniques discussed in previous sections. However, such bounds are typically very conservative. Some loose intuition for this is that projecting the multi-dimensional process $\{X_t\}_{t\geq 0}$ onto a one-dimensional space through V results in a large loss of information.

Strategies for developing direct bounds on the mean-squared error of $\tilde{f}_m$ are presented using two different approaches in [63] and [58]. In the first, regeneration times are exploited, whereas in the second some ideas from Geometry and optimal transport are used to construct a bound conditional on a positive Ricci curvature condition for the chain. Both approaches seem to hold some promise, and have been further analysed and extended (e.g. [113, 37]).

3.3 Diffusion processes

Although we are primarily interested in discrete-time stochastic processes, sometimes these can be effectively constructed by first considering continuous-time processes. Often there is more structure to exploit in the continuous case, provided by some form of differential calculus. Here we discuss some such processes which can be used to build Markov chain Monte Carlo methods.

We do not seek here to be as thorough and pedagogical as in the section for discrete-time chains, which are our primary focus. Most of the work for this thesis is concerned with analysis of chains inspired by processes, not the processes themselves. For this reason, this section is confined to

reviewing some key results which are needed. For a thorough introduction to continuous-time processes, see [67, 89].

We define a continuous-time Markov process as a collection of random variables $(X_t)_{t\geq 0}$, indexed by some continuous parameter $t \geq 0$. in many cases the index is called time, though processes have also been studied in Statistics which evolve across space (often referred to as *random fields*) and other domains. For any fixed t, X_t is a random variable.

In the continuous-time setting, the *Markov* property is most easily stated using the concept of a *filtration*, an increasing family of σ-algebras $\{\mathscr{F}_t\}_{t\geq 0}$ for which $\mathscr{F}_t$ contains the 'history' of the process up until time t. The Markov property can then be defined as

$$\mathbb{P}[X_{t+h} \in A|\mathscr{F}_t] = \mathbb{P}[X_{t+h} \in A|X_t].$$

A more intuitive way to think of this for the less mathematically inclined is that for any collection of times $\{t_i\}_{i=1}^n$ with each $t_i \leq t_0$, any $h > 0$ and any $A \in \mathscr{B}$

$$\mathbb{P}[X_{t_0+h} \in A|X_{t_0} = x_{t_0}, X_{t_1} = x_{t_1}, ..., X_{t_n} = x_{t_n}] = \mathbb{P}[X_{t_0+h} \in A|X_{t_0} = x_{t_0}].$$

We will primarily discuss the class of Markov processes known as *diffusions*, those for which 'sample paths' $(X_t(\omega))_{t\geq 0}$ are continuous with probability one. What is meant by this is that if we consider the set of possible paths as deterministic functions $f_\omega : [0,\infty) \to \mathbf{X}$ which map $t \to X_t(\omega)$, the set of outcomes $\omega \in \Omega$ for which f_ω is not a continuous function occurs with probability zero. The remainder of this section is adapted from [71] (which is the author's own work).

We focus on the class of time-homogeneous Itô diffusions, whose dynamics are governed by a stochastic differential equation of the form:

$$dX_t = b(X_t)dt + \sigma(X_t)dB_t, \;\; X_0 = x_0, \tag{3.24}$$

where $(B_t)_{t\geq 0}$ is a standard Brownian motion. To unpack this slightly, by *Brownian motion* we mean a process with $B_0 = 0$, with independent increments, for which $B_t \sim N(0,tI)$. $(B_t)_{t\geq 0}$ is also continuous with probability one, but is nowhere differentiable (an interesting discussion on this point is given in Section 13.2 of [44]). The drift vector b and volatility matrix σ in (3.24) are usually assumed to be Lipschitz[6] continuous functions, for reasons discussed in Section 5.2 of [89], however

[6] A function $f : X \to Y$ which maps from one metric space (X, d_X) to another (Y, d_Y) is called *Lipschitz* if there is a real constant $K < \infty$ such that $d_Y(f(x), f(y)) \leq K d_X(x,y)$ for any $x, y \in X$.

this is not always necessary, and we give some examples which are not Lipschitz in Chapter 5. Under this assumption, however, and noting that $\mathbb{E}[B_{t+h} - B_t | X_t = x_t] = 0$ for any $h \geq 0$, informally we can see that

$$\mathbb{E}[X_{t+h} - X_t | X_t = x_t] = b(x_t)h + o(h),$$

implying that the drift dictates how the mean of the process changes over a small time interval. In addition, if we define the process $(M_t)_{t\geq 0}$ through the relation

$$M_t = X_t - \int_0^t b(X_s)ds,$$

then we have

$$\mathbb{E}[(M_{t+h} - M_t)(M_{t+h} - M_t)^T | M_t = m_t, X_t = x_t] = \sigma(x_t)\sigma(x_t)^T h + o(h),$$

giving the stochastic part of the relationship between X_{t+h} and X_t for small enough h. See e.g. Section 5.1 of [111].

Although (3.24) is often a suitable description of an Itô diffusion, we can characterise them in several different ways. As in the discrete time case, a diffusion can be described through a transition kernel $P^t(x_0, \cdot)$. Typically, however, the form of $P^t(x_0, \cdot)$ is unknown, though we can write the expectation and variance of $X_t \sim P^t(x_0, \cdot)$ via the integral equations

$$\mathbb{E}[X_t | X_0 = x_0] = x_0 + \mathbb{E}\left[\int_0^t b(X_s)ds\right],$$
$$\mathbb{E}[(X_t - \mathbb{E}[X_t|X_0 = x_0])(X_t - \mathbb{E}[X_t|X_0 = x_0])^T | X_0 = x_0] = \mathbb{E}\left[\int_0^t \sigma(X_s)\sigma(X_s)^T ds\right],$$

where the second of these is as a result of the Itô isometry (see e.g. page 29 of [89]).

Another (often more tractable) way to characterise a diffusion process is through an *infinitessimal generator*, $\mathscr{A}$, which describes how functions of the process are expected to evolve. We define this partial differential operator through its action on a function $f \in C^2(\mathbf{X})$ as[7]

$$\mathscr{A}f = \lim_{h\to 0} \frac{\mathbb{E}[f(X_{t+h})|X_t = x_t] - f(x_t)}{h},$$

though $\mathscr{A}$ can be associated with the drift and volatility of $(X_t)_{t\geq 0}$ by the relation

$$\mathscr{A}f(x) = \sum_i b_i(x)\frac{\partial f}{\partial x_i}(x) + \frac{1}{2}\sum_{i,j} A_{ij}(x)\frac{\partial^2 f}{\partial x_i x_j}(x), \tag{3.25}$$

[7] $C^2(\mathbf{X})$ is the set of functions $f : \mathbf{X} \to \mathbb{R}$ with continuous first and second partial derivatives.

where $A_{ij}(x)$ denotes the component in row i and column j of $\sigma(x)\sigma(x)^T$ (see Section 7.3 of [89]).

In the case of an Itô diffusion, provided $A(x)$ is positive definite for all x then $P^t(x_0,\cdot)$ admits a density $p_t(x|x_0)$, which, in fact, varies smoothly as a function of t. The Fokker–Planck equation[8] describes the variation in terms of the drift and volatility and is given by

$$\frac{\partial}{\partial t}p_t(x|x_0) = -\sum_i \frac{\partial}{\partial x_i}[b_i(x)p_t(x|x_0)] + \frac{1}{2}\sum_{i,j}\frac{\partial^2}{\partial x_i \partial x_j}[A_{ij}(x)p_t(x|x_0)]. \tag{3.26}$$

A natural question is whether a diffusion has an invariant measure $\pi(\cdot)$, and whether as $t \to \infty$

$$\|P^t(x_0,\cdot) - \pi(\cdot)\|_{TV} \to 0,$$

for any $x_0 \in \mathbf{X}$. Again, similarly to the discrete time case, we require the diffusion to be positive Harris recurrent with $\pi(\cdot)$ as an invariant distribution, where here positive and Harris are defined analogously to in the discrete case. In addition to this, there is a topological constraint that all compact sets must be small for some skeleton chain. See [80] for details. Equation (3.26) actually provides a means of finding $\pi(\cdot)$, given b and σ, highlighting that the added structure offered by some continuous-time processes can be of use. Setting the left-hand side of (3.26) to zero gives

$$\sum_i \frac{\partial}{\partial x_i}[b_i(x)\pi(x)] = \frac{1}{2}\sum_{i,j}\frac{\partial^2}{\partial x_i x_j}[A_{ij}(x)\pi(x)], \tag{3.27}$$

which can be solved to find $\pi(\cdot)$.

In the case $t \in \mathbb{R}_{>0}$, φ-irreducibility can be defined in a similar way to $t \in \mathbb{Z}_{>0}$. We say a process $(X_t)_{t\geq 0}$ is φ-irreducible if for any $x \in \mathbf{X}$, there exists a $t = t(x,A) > 0$ for which

$$\varphi(A) > 0 \Rightarrow P^t(x,A) > 0, \ \ \forall A \in \mathscr{B}.$$

We can equivalently say $\varphi(A) > 0 \Rightarrow \mathbb{E}_x[\tau_A] > 0$.

The continuous-time analogue to geometric ergodicity is not surprisingly referred to using the continuous equivalent of a geometric random variable. We say that a π-irreducible process $(X_t)_{t\geq 0}$ with $X_0 = x_0$ is *exponentially ergodic* if

$$\|P^t(x_0,\cdot) - \pi(\cdot)\|_{TV} \leq M(x_0)r^t, \tag{3.28}$$

for some $r < 1$ and $M : \mathbf{X} \to [0,\infty)$ [80].

[8] Also known as the Kolmogorov forward equation.

Again, drift and minorisation conditions can be exploited to establish (3.28). Using the generator characterisation, the equivalent condition to (3.21) is

$$\mathscr{A}V(x) \leq -cV(x) + b\mathbb{1}_C(x), \tag{3.29}$$

for some small set C, for any $x \in \mathbf{X}$, where $c > 0$, $b < \infty$. In some sense the continuous analogue is more accessible than (3.21). Through the generator characterisation, we have actually defined our process via how we *expect* it to evolve. So no further integrals are required in (3.29), as opposed to (3.21). Note that in the continuous case we refer to a set $C \subset \mathbf{X}$ as small if for any $A \in \mathscr{B}$ there exists $t \geq 0$ such that.

$$P^t(x_0, A) \geq \varepsilon\psi(A), \ \forall x_0 \in C,$$

for some probability measure $\psi(\cdot)$. This is analogous to (3.20).

Chapter 4

Markov chain Monte Carlo methods

In this chapter we review the Metropolis–Hastings algorithm, and some popular delineations. Although this is not the only way to construct a measure-preserving Markov chain, it is both easy to do and very general, making it the 'go to' choice in Markov chain Monte Carlo [29].

The basic premise of the method is to construct a transition kernel P such that the *detailed balance* equations

$$\pi(dx)P(x,dy) = \pi(dy)P(y,dx) \tag{4.1}$$

are satisfied for the distribution of interest $\pi(\cdot)$, for any $x, y \in \mathbf{X}$. Integrating over x gives

$$\int \pi(dx)P(x,dy) = \int \pi(dy)P(y,dx) = \pi(dy)\int P(y,dx) = \pi(dy),$$

showing that (4.1) is a sufficient (but not necessary) condition for $\pi(\cdot)$ to be stationary for $\{X_t\}_{t\geq 0}$. Readers unsatified with the formal infinitessimal notation used here can consult Section 20.1.2 of [81] for a more detailed treatment. If we can also establish π-irreducibility, aperiodicity and Harris recurrence, then the chain will be ergodic, with $\pi(\cdot)$ as the unique invariant distribution. In words, (4.1) states that the probability of the chain being in a set $A \in \mathscr{B}$ and moving to $B \in \mathscr{B}$ is the same as that of being at B and moving to A, for any $A, B \in \mathscr{B}$. Markov chains that satisfy (4.1) are called *reversible*, because if the chain is at stationarity then

$$\mathbb{P}[X_n \in A, X_{n+1} \in B] = \mathbb{P}[X_n \in B, X_{n+1} \in A],$$

so that it is not possible to identify whether time is moving forwards or backwards for the chain.

4.1 Metropolis–Hastings

In a similar vain to the rejection and importance sampling methods, the Metropolis–Hastings algorithm can be viewed as a *re-sampling* approach. A Markov chain is constructed by first drawing some 'proposed' next position in the chain from some candidate transition kernel Q, and then using some accept/reject mechanism to ensure that the full transition kernel P satisfies (4.1). The full algorithm is given below (with $a \wedge b$ denoting the minimum of a and b).

Algorithm 1 Metropolis–Hastings, single iteration.

Require: x_{i-1}

Draw $X' \sim Q(x_{i-1}, \cdot)$

Draw $Z \sim U[0,1]$

Set $\alpha(x_{i-1}, x') \leftarrow 1 \wedge \frac{\pi(x')q(x_{i-1}|x')}{\pi(x_{i-1})q(x'|x_{i-1})}$

if $z < \alpha(x_{i-1}, x')$ **then**

Set $x_i \leftarrow x'$

else

Set $x_i \leftarrow x_{i-1}$

end if

The full transition kernel for a Markov chain constructed using the Metropolis–Hastings method is

$$P(x,A) = \int_A \alpha(x,y)Q(x,dy) + r(x)\delta_x(A) \tag{4.2}$$

for any $A \in \mathscr{B}$, where

$$r(x) = 1 - \int \alpha(x,y)Q(x,dy)$$

is the average probability that a proposed moved from $Q(x,\cdot)$ will be rejected.

Proposition 4.1. *A Markov chain $\{X_t\}_{t\geq 0}$ produced by the Metropolis–Hastings algorithm has $\pi(\cdot)$ as an invariant distribution.*

Proof: We first show that $\alpha(x,y)Q(x,dy)$ satisfies detailed balance, and then establish the result. Assuming $\pi(\cdot)$ and $Q(x,\cdot)$ admit densities $\pi(x)$ and $q(y|x)$ (which is always possible by constructing

the reference measure $\mathbf{m}(\cdot) = \pi(\cdot) + Q(x,\cdot)$), we can write

$$\begin{aligned}\pi(dx)\alpha(x,y)Q(x,dy) &= \pi(x)q(y|x) \wedge \pi(y)q(x|y)\mathbf{m}(dx)\mathbf{m}(dy) \\ &= \pi(dy)\alpha(y,x)Q(y,dx).\end{aligned}$$

To show that $\pi(\cdot)$ is invariant for P, note that

$$\begin{aligned}\int \pi(dx)P(x,A) &= \int \pi(dx)\left[r(x)\delta_x(A) + \int_A \alpha(x,y)Q(x,dy)\right], \\ &= \int_A r(x)\pi(dx) + \int_{y\in A}\int \alpha(x,y)\pi(dx)Q(x,dy), \\ &= \int_A r(x)\pi(dx) + \int_A \left[\int \alpha(y,x)Q(y,dx)\right]\pi(dy), \\ &= \int_A r(x)\pi(dx) + \int_A (1-r(y))\pi(dy) = \pi(A),\end{aligned}$$

as required. ■

Provided that the acceptance probability is less than one in some region of $\mathbf{X}$, then the chain produced will be aperiodic, as there will be a non-zero probability of remaining in the same part of the state space. Of course, whether or not the chain is ψ-irredicible for some $\pi(\cdot) \ll \psi(\cdot)$, Harris recurrent and geometrically ergodic will depend crucially on the choice of Q. We now discuss some typical options.

4.1.1 Independence sampler

Perhaps the simplest choice for Q is something which is independent of the current position, meaning $Q(x,\cdot) = q(\cdot)$ for any $x \in \mathbf{X}$. In this instance the acceptance probability reduces to

$$\alpha(x,y) = \frac{\pi(y)q(x)}{\pi(x)q(y)}.$$

Note that X_{i+1} will still depend on X_i through α. Intuitively, the best choice for $q(\cdot)$ is $q(\cdot) \approx \pi(\cdot)$, so that $\alpha \approx 1$ and the algorithm *almost* produces independent samples from $\pi(\cdot)$.

Unfortunately, the independence sampler suffers from the same drawbacks as other re-sampling methods, in that finding a distribution $q(\cdot)$ which globally approximates $\pi(\cdot)$ is extremely difficult in high dimensions. More formally, ergodicity results for the independence sampler highlight exactly when the algorithm should and should not be used.

Theorem 4.2. *If the candidate distribution $q(\cdot)$ satisfies the 'heavy-tail rule'*

$$\frac{q(x)}{\pi(x)} \geq \delta, \ \ \forall x \in \mathbf{X}, \tag{4.3}$$

for some $\delta > 0$, then the independence sampler produces a Markov chain which is uniformly ergodic. If (4.3) is not satisfied, then the chain will not even be geometrically ergodic.

Proof: This is Theorem 2.1 in [78].

In practice (4.3) is very difficult to establish, as noted by Johnson & Geyer [54], making the independence sampler difficult to use with confidence, particularly for high-dimensional models.

4.1.2 Random Walk Metropolis

Another extremely simple choice for Q is one in which the resulting transition density $q(y|x)$ satisfies

$$q(y|x) = q(|y-x|),$$

meaning the proposal density is symmetric about x. A simple example is $Q(x,\cdot) = N(x, \lambda^2\Sigma)$ for some $\lambda^2 > 0$, where Σ is either taken simply as the identity, or chosen to match the correlation structure of $\pi(\cdot)$. In this symmetric case, α reduces to

$$\alpha(x,y) = 1 \wedge \frac{\pi(y)}{\pi(x)} \tag{4.4}$$

which has the clean interpretation that proposals for which the target density is larger will be accepted. More generally, any choice of Q for which α reduces to (4.4) is called a *Metropolis* algorithm [48]. The Random Walk Metropolis (RWM) is a special case.

Much theoretical study has been dedicated to the RWM. It has been shown that the optimal acceptance rate for proposals tends to 0.234 as the dimension n of $\mathbf{X}$ tends to ∞ for a wide class of targets [101, 120]. The intuition for an optimal acceptance rate is to find the right balance between proposing moves which are far from the current point in the chain and ensuring that these moves will be accepted a reasonable proportion of the time, so as to minimise the asymptotic variance (3.17). If a proposal y is 'close' to the current point x, then $\pi(y)/\pi(x) \approx 1$, so the acceptance probability will be high, but $\mathrm{Corr}_\pi[f(X_{i+1}), f(X_i)]$ will typically be close to 1, increasing the variance of the estimator. However, if y is far away from x, it could easily be that $\pi(y)/\pi(x) \approx 0$, meaning the chain stays put

and $\text{Corr}_\pi[f(X_{i+1}), f(X_i)] = 1$, which is clearly undesirable. Random walk proposals are sometimes referred to as 'blind', as no information about $\pi(\cdot)$ is used when generating proposals, so typically very large moves will result in a very low chance of acceptance.

Several authors have also shown that for certain classes of $\pi(\cdot)$, the tuning parameter λ (highlighted above in the Gaussian case of Q) should be chosen such that $\lambda^2 \propto n^{-1}$, so that $\alpha \not\to 0$ as $n \to \infty$ [101]. Because of this, we say that algorithm efficiency 'scales' $O(n^{-1})$ as the dimension of $\mathbf{X}$ increases. Note that this compares favourably with both numerical and traditional re-sampling methods.

Ergodicity results for a Markov chain constructed using the RWM algorithm also exist [78, 110]. At least exponentially-light tails are a necessity for $\pi(x)$ for geometric ergodicity (defined precisely in Subsection 4.2). In higher dimensions additional conditions are required [110]. These ergodicity properties are discussed in much more detail in Subsection 4.2. We demonstrate with a simple example why heavy-tailed forms of $\pi(x)$ pose difficulties here (where $\pi(x) \to 0$ at a slower than exponential rate).

Example 4.3. *Take $\pi(x) \propto 1/(1+x^2)$, so that $\pi(\cdot)$ is a Cauchy distribution. Then if $Y \sim N(x, \lambda^2)$, the ratio $\pi(y)/\pi(x) = (1+x^2)/(1+y^2) \to 1$ as $|x| \to \infty$. Therefore, if x_0 is far away from zero, the Markov chain will dissolve into a random walk, with almost every proposal being accepted.*

Of course, a one-dimensional random walk is an example of a *null* recurrent Markov chain, which does not have a finite invariant measure (see e.g. Section 21.4 of [66]). It should be noted that starting the chain from near zero can also cause problems in this example, as the tails of the distribution may not be explored suitably quickly. See [100] for more details here.

4.1.3 Metropolis-adjusted Langevin algorithm

We have already referred to random walk proposals as 'blind', as no information about $\pi(\cdot)$ is used to generate them. Intuitively, it would seem more sensible to construct Q in a way such that $\pi Q(\cdot) \approx \pi(\cdot)$, so that the majority of proposals will be accepted. One way to construct such a candidate kernel would be to base it on a diffusion which has $\pi(\cdot)$ as a limiting distribution, and then use the Metropolis–Hastings step simply to correct for the error introduced by numerically simulating the process.

Given the Fokker–Planck equation (3.26), our goal therefore becomes clear: find drift and volatility

terms such that the resulting dynamics describe a diffusion which converges to some user-defined invariant distribution, $\pi(\cdot)$. This process can then be used as a basis for choosing Q in a Metropolis–Hastings algorithm. The Langevin diffusion, first used to describe the dynamics of molecular systems [23], is such a process, given by the solution to the stochastic differential equation:

$$dX_t = \frac{1}{2}\nabla \log \pi(X_t)dt + dB_t, \;\; X_0 = x_0. \tag{4.5}$$

Since the volatility terms are $A_{ij}(x) = \mathbb{1}_{\{i=j\}}$, it is clear that

$$\frac{1}{2}\frac{\partial}{\partial x_i}[\log \pi(x)]\pi(x) = \frac{1}{2}\frac{\partial}{\partial x_i}\pi(x), \;\; \forall i, \tag{4.6}$$

which is a sufficient condition for Equation (3.26) to hold. Therefore, for any case in which $\pi(x)$ is suitably regular (so that $\nabla \log \pi(x)$ is well-defined and the derivatives in Equation (3.26) exist), we can use (4.5) to construct a diffusion which has invariant distribution $\pi(\cdot)$.

Roberts and Tweedie [109] give sufficient conditions on $\pi(\cdot)$ under which a diffusion $(X_t)_{t\geq 0}$ with dynamics given by Equation (4.5) will be ergodic, meaning

$$\|\boldsymbol{P}^t(x_0,\cdot) - \pi(\cdot)\|_{TV} \to 0 \tag{4.7}$$

as $t \to \infty$, for any $x_0 \in \mathbf{X}$. They are straightforwardly satisfied by many statistical models.

We can use Langevin diffusions as a basis for MCMC in many ways, but a popular variant is known as the Metropolis-adjusted Langevin algorithm (MALA), in which $Q(x,\cdot)$ is constructed through an Euler–Maruyama discretisation of (4.5) and used as a candidate kernel in a Metropolis–Hastings algorithm. The resulting proposal is:

$$Q(x,\cdot) = N\left(x + \frac{\lambda^2}{2}\nabla \log \pi(x), \lambda^2 I\right), \tag{4.8}$$

where λ is again a tuning parameter.

Before we discuss the theoretical properties of the approach, we first offer some intuition for the dynamics. From Equation (4.8), it can be seen that Langevin-type proposals comprise a deterministic shift towards the local mode of $\pi(x)$, combined with some random additive Gaussian noise, with variance λ^2 for each component. The relative weights of the deterministic and random parts are fixed, given as they are by the parameter λ. Typically, if $\lambda \gg \lambda^2$, then the random part of the proposal will dominate and *vice versa* in the opposite case, though this also depends on the form of $\nabla \log \pi(x)$ [109].

Again, since this is a Metropolis–Hastings method, choosing λ is a balance between proposing large enough jumps and ensuring that a reasonable proportion are accepted. It has been shown that in the limit (as $n \to \infty$), the optimal acceptance rate for the algorithm is 0.574 [104] for forms of $\pi(\cdot)$ which either have independent and identically distributed components or whose components only differ by some scaling factor [104]. In these cases, as $n \to \infty$, the parameter λ^2 must be chosen $\propto n^{-1/3}$, so we say that algorithm efficiency scales $O(n^{-1/3})$. Note that these results compare favourably with the $O(n^{-1})$ scaling of the Random Walk Metropolis.

Convergence properties of the method have also been established. Roberts and Tweedie [109] highlight some cases in which MALA is either geometrically ergodic or not. Typically, results are based on the tail behaviour of $\pi(x)$. If these tails are heavier than exponential, then the method is typically not geometrically ergodic and similarly if the tails are lighter than Gaussian. However, in the in-between case, the converse is true. We again offer two simple examples for intuition here.

Example 4.4. *Take $\pi(x) \propto 1/(1+x^2)$ as in the previous example. Then, $\nabla \log \pi(x) = -2x/(1+x^2)^2 \to 0$ as $|x| \to \infty$. Therefore, if x_0 is far away from zero, then the MALA will be approximately equal to the RWM algorithm, and so will also dissolve into a random walk.*

Example 4.5. *Take $\pi(x) \propto e^{-x^4}$. Then, $\nabla \log \pi(x) = -4x^3$ and $X' \sim N(x - 2\lambda^2 x^3, \lambda^2)$. Therefore, for any fixed λ, there exists $c > 0$, such that, for $|x| > c$, we have $|x - 2\lambda^2 x^3| >> |x|$, suggesting that MALA proposals will quickly spiral further and further away from any neighbourhood of zero, and hence nearly all will be rejected.*

For cases where there is a strong correlation between elements of x or each element has a different marginal variance, the MALA can also be 'pre-conditioned' in a similar way to the RWM, so that the covariance structure of proposals more accurately reflects that of $\pi(x)$ [108]. In this case, proposals take the form

$$Q(x,\cdot) = N\left(x + \frac{\lambda^2}{2}\Sigma\nabla \log \pi(x), \lambda^2\Sigma\right). \tag{4.9}$$

It can be shown that provided Σ is a constant matrix, $\pi(x)$ is still the invariant distribution for the diffusion on which Equation (4.9) is based [130].

4.1.4 Hamiltonian Monte Carlo

The next algorithm we introduce is descended from the Physics literature [35]. In Hamiltonian Monte Carlo (HMC), the usual state space $\mathbf{X}$ is doubled in size, with the introduction of auxiliary 'momentum' variables to accompany the 'position' variables x in a $2n$-dimensional physical system. If we define the *Hamiltonian* function

$$H(x,p) = -\log \pi(x) + \frac{1}{2}\log|G(x)| + \frac{1}{2}p^T G^{-1}(x)p,$$

then we can construct a probability density on this augmented space as $f(x,p) \propto e^{-H(x,p)}$, with marginals $X \sim \pi(\cdot)$ and $p \sim N(0, G(x))$. Continuing the physical analogy, $H(x,p)$ is the total energy in the system, with $U(x) = -\log \pi(x)$ the potential energy and $K(x,p) = p^T G^{-1}(x)p/2$ the kinetic.

Hamiltonian dynamics are a way to evolve (x,p) in such a way that $H(x,p)$ remains constant (an energy conserving system). Trivially

$$\frac{d}{dt}H(x,p) = \sum_i \left(\frac{\partial H}{\partial x_i}\frac{dx_i}{dt} + \frac{\partial H}{\partial p_i}\frac{dp_i}{dt} \right),$$

so a simple way to ensure $dH/dt = 0$ is to evolve (x,p) using the dynamics

$$\frac{dx_i}{dt} = \frac{\partial H}{\partial p_i}, \quad \frac{dp_i}{dt} = -\frac{\partial H}{\partial x_i}.$$

We define the action of *Hamiltonian flow* for t units of time via the map $(x,p) \to \xi_t(x,p)$. Clearly $H(\xi_t(x,p)) = H(x,p)$ for any t.

The map is measure-preserving, but its repeated application would result in flowing along a density contour rather than exploring the entire state space $\mathbf{X} \times \mathbf{X}$, so the Hamiltonian Monte Carlo algorithm intersperses 'momentum refreshment' steps. The full construction is:

Algorithm 2 Idealised Hamiltonian Monte Carlo, single iteration.

Require: x_{i-1}, $T \in \mathbb{R}^+$

Draw $p|X = x_{i-1} \sim N(0, G(x_{i-1}))$

Set $(x_i, p_i) \leftarrow \xi_T(x_{i-1}, p)$

Heuristic discussion of ergodic properties is given in [84]. We provide a more formal treatment in Chapter 7. The difficulty in practice is that Hamiltonian flow is often analytically intractable. Fortunately, there are numerical schemes for Hamiltonian dynamics which, while not measure-preserving,

are still both volume-preserving, and produce maps which can be constructed so as to be reversible [84]. If the Hamiltonian is separable (meaning $K(x,p) = K(p)$ here), which in the above scheme means $G(x) = M$, then one such numerical scheme is the leapfrog method, or Störmer–Verlet integrator [65]. The three steps comprising a single iteration for time step ε is

$$\begin{aligned} p_{t+\varepsilon/2} &= p_t + \varepsilon \nabla \log \pi(x_t)/2, \\ x_{t+\varepsilon} &= x_t + \varepsilon M^{-1} p_{t+\varepsilon/2}, \\ p_{t+\varepsilon} &= p_{t+\varepsilon/2} + \varepsilon \nabla \log \pi(x_{t+\varepsilon})/2. \end{aligned}$$

This flow does not in general preserve the Hamiltonian $H(x,p)$, but does have unit determinant (volume preservation), which implies the density for the resulting random variable $\eta_t^*(x,p)$ is still $\propto e^{-H(\eta_t^*(x,p))}$ for any $t \in \mathbb{R}$. We can construct a flow which is also *reversible* [84] by running the leapfrog scheme for $T = L\varepsilon$ time units and then negating the momentum. We denote this numerical flow plus negation $(x,p) \to \eta_T(x,p)$. We also write $\eta_T^x(x,p)$ for the x-coordinate of the resulting map, and the same for p.

To see that the leapfrog integrator is volume preserving, we show that the determinant of each of the mappings described above is one. Here we write for the first step $\psi_1(x_t, p_t) = (x_t, p_{t+\varepsilon/2})$, and similarly for the second, and $\psi_1^x(x_t,p_t) = x_t$, $\psi_1^p(x_t,p_t) = p_{t+\varepsilon/2}$ as the projections onto each coordinate. With this notation we have

$$(x_{t+\varepsilon}, p_{t+\varepsilon}) = \psi_1 \circ \psi_2 \circ \psi_1(x_t, p_t).$$

Next note that ψ_1^x is simply the identity, while ψ_1^p is the current momentum plus some terms which only depend on the position coordinate. So denoting $\partial_x := \partial/\partial_x$, the jacobian for this mapping is

$$J(\psi_1) = \begin{pmatrix} 1 & 0 \\ \partial_x \psi_1^p & 1 \end{pmatrix},$$

which has unit determinant. A similar argument holds for ψ_2. Such mapping are called *shear transformations*, which are known to preserve volume (e.g. [84]).

To see that the flow induced by the leapfrog integrator is reversible, note that the composition of maps gives (taking $M = I$ without loss of generality):

$$x_{t+\varepsilon} = x_t + \varepsilon^2 \nabla \log \pi(x_t)/2 + \varepsilon p_t, \tag{4.10}$$

$$p_{t+\varepsilon} = p_t + \varepsilon \nabla \log \pi(x_t)/2 + \varepsilon \nabla \log \pi(x_{t+\varepsilon})/2, \tag{4.11}$$

Using these means that considering the flow initialised from $(x_{t+\varepsilon}, -p_{t+\varepsilon})$ we have

$$\begin{aligned}
\eta_\varepsilon^x(x_{t+\varepsilon}, -p_{t+\varepsilon}) &= x_{t+\varepsilon} + \varepsilon^2 \nabla \log \pi(x_{t+\varepsilon}) - \varepsilon p_{t+\varepsilon}, \\
&= x_{t+\varepsilon} + \varepsilon^2 \nabla \log \pi(x_{t+\varepsilon})/2 - \varepsilon \left(p_t + \varepsilon \nabla \log \pi(x_t)/2 + \varepsilon \nabla \log \pi(x_{t+\varepsilon})/2\right), \\
&= x_t + \varepsilon^2 \nabla \log \pi(x_t)/2 + \varepsilon p_t - \varepsilon(p_t + \varepsilon \nabla \log \pi(x_t)/2), \\
&= x_t.
\end{aligned}$$

Similarly we have

$$\begin{aligned}
\eta_t^p(x_{t+\varepsilon}, -p_{t+\varepsilon}) &= (-1) \times (-p_{t+\varepsilon} + \varepsilon \nabla \log \pi(x_{t+\varepsilon})/2 + \varepsilon \nabla \log \pi(\eta_\varepsilon^x(x_{t+\varepsilon}, -p_{t+\varepsilon}))/2), \\
&= (-1) \times (-p_{t+\varepsilon} + \varepsilon \nabla \log \pi(x_{t+\varepsilon})/2 + \varepsilon \nabla \log \pi(x_t)/2),
\end{aligned}$$

which, when substituting the right-hand side of (4.11) for $p_{t+\varepsilon}$ and simplifying, gives p_t.

In the case where the Hamiltonian is *not* separable, explicit symplectic integrators are no longer available [43]. A possible scheme in this case is given in [43].

To correct for bias in the resulting MCMC estimators, after each numerical flow step a Metropolis accept-reject step is used (note that the full Hastings acceptance rate is not needed as the mapping η_t is reversible $\forall t \in \mathbb{R}$). It is shown in [35] that the resulting scheme shown below is π-invariant. A more general proof of this is given in [125], which holds for any deterministic proposal $y = g(x)$ for which g is an *involution*, meaning $g \circ g(x) = x$. It is straightforward to see that including the momentum negation step makes the leapfrog flow $\eta_t(x, p)$ an involution for any fixed t.

Algorithm 3 Hamiltonian Monte Carlo, single iteration.

Require: x_{i-1}, $\varepsilon \geq 0$, $L \in \mathbb{N}$

Draw $p \sim N(0, M)$

Set $T \leftarrow L\varepsilon$, propose $(x', p') = \eta_T(x_{i-1}, p)$

Draw $Z \sim U[0,1]$

Set $\alpha \leftarrow 1 \wedge e^{-\delta H}$, where $\delta H = H(x', p') - H(x_{i-1}, p)$

if $z < \alpha$ **then**

Set $x_i \leftarrow x'$

else

Set $x_i \leftarrow x_{i-1}$

end if

Despite numerous authors noting the strong empirical performance of the method [84, 43], much less has been established regarding the theoretical properties of HMC [29]. Beskos et al. [6] show that a lower bound for the optimal acceptance rate α is 0.651 as $n \to \infty$, and that algorithm efficiency scales $O(n^{-1/4})$ for certain classes of targets, which compares favourably to all other methods presented here. Betancourt, Byrne and Girolami [10] establish an upper bound on this rate of 0.9, and also broaden the class of targets for which it applies. Difficulty in choosing the integration time T and mass matrix $G(x)$, together with the computational overhead of the numerical integration scheme, have however been noted as possible stumbling blocks to wider implementation [84, 96], though some recent suggestions have been made for each of these [43, 49, 123, 9]. In the next sections in particular we review the geometric intuition behind certain choices of $G(x)$.

The question of irreducibility is subtle here: while the method is generally assumed to satisfy this property with the exception of some pathological special cases (discussed further in Chapter 7), the result has only been proven in the case where $\pi(x) \geq c > 0$ for all $x \in \mathbf{X}$ [19]. The general question (in the Lebesgue case) is whether for any $B \in \mathscr{B}$ with $\mu^L(B) > 0$ then

$$Q(x,B) = \int \mathbb{1}_{B\times\mathbf{X}}(\eta_T(x,p))\mu^G(dp) > 0,$$

where $\mu(\cdot)$ denotes a standard Gaussian measure. We give a more thorough treatment of this issue in Chapter 7.

4.2 Geometric ergodicity of Metropolis–Hastings methods

Roberts & Tweedie [110] simplified the matter of establishing geometric ergodicity for a Markov chain by showing that if all compact sets of $\mathbf{X}$ are small, then we need not explicitly find a small set C, but instead can show that there is a Lyapunov function V for which

$$\limsup_{|x|\to\infty} \int \frac{V(y)}{V(x)} P(x,dy) < 1. \tag{4.12}$$

In fact (4.12) is both necessary and sufficient for geometric ergodicity [110]. In the case where P is a Metropolis–Hastings kernel, then we have

$$\int \frac{V(y)}{V(x)} P(x,dy) = \int \frac{V(y)}{V(x)} \alpha(x,y) Q(x,dy) + r(x) = \int \left[\frac{V(y)}{V(x)} - 1\right] \alpha(x,y) Q(x,dy) + 1. \tag{4.13}$$

This means that an equivalent condition to (4.12) is

$$\limsup_{|x|\to\infty} \int \left[\frac{V(y)}{V(x)} - 1\right] \alpha(x,y) Q(x,dy) < 0, \tag{4.14}$$

where Q is the proposal kernel and α the acceptance rate. The authors also note that a sufficient condition for *lack* of geometric ergodicity is

$$\operatorname*{ess\,sup}_{|x|\to\infty} r(x) = 1. \tag{4.15}$$

Intuitively this impies that the chain is likely to get 'stuck' for large periods. In the context of Metropolis–Hastings, the authors also established sufficient conditions for all compact sets of $\mathbf{X}$ to be small in terms of $\pi(x)$ and $q(y|x)$, where $Q(x,dy) = q(y|x)dx$.

Theorem 4.6. *(Roberts & Tweedie). Suppose that $\pi(x)$ is bounded away from 0 and ∞ on compact sets, and there exists $\delta_q > 0$ and $\varepsilon_q > 0$ such that, for every x*

$$|x-y| \leq \delta_q \Rightarrow q(y|x) \geq \varepsilon_q.$$

Then the chain with kernel (4.2) is μ^L-irreducible and aperiodic, and every nonempty compact set is small.

Jarner & Tweedie [52] introduced a necessary condition for geometric ergodicity through a *tightness* condition.

Theorem 4.7. *(Jarner & Tweedie). If for any $\xi > 0$ there is a $\delta > 0$ such that for all $x \in \mathbf{X}$*

$$P(x, B_\delta(x)) > 1 - \xi,$$

where $B_\delta(x) := \{y \in \mathbf{X} : d(x,y) < \delta\}$, then P can be geometrically ergodic only *in the case where for some $s > 0$*

$$\int e^{s|x|}\pi(dx) < \infty.$$

The result highlights that when $\pi(\cdot)$ is heavy-tailed the chain must be able to make very large moves and still be capable of returning to the centre quickly for the geometric total variation distance bound (3.18) to hold. In the Metropolis–Hastings case it is straightforward to see that

$$Q(x, B_\delta(x)) > 1 - \xi \Rightarrow P(x, B_\delta(x)) > 1 - \xi, \tag{4.16}$$

which is a useful approach to establishing lack of geometric ergodicity in the heavy-tailed case.

4.2.1 Random Walk Metropolis in one dimension

In this section a 'textbook style' proof is provided of geometric ergodicity for the Random Walk Metropolis in one dimension. The result was first published in [78]. Before establishing a positive result, we first note that the tightness condition (4.16) holds here, so it is straightforward to see that the algorithm can *only* produce a geometricaly ergodic Markov chain if there is an $s > 0$ such that $\mathbb{E}_\pi[e^{s|x|}] < \infty$. In one dimension such a restriction on $\pi(\cdot)$ is known as *tail log-concavity*. Specifically, the density $\pi(x)$ is called *log-concave in the tails* if for some $x_0 > 0$, $a > 0$ and all $y \geq x \geq x_0$ we have

$$\pi(y)/\pi(x) \leq e^{-a(y-x)},$$

and a similar condition holds in the negative tail.

The skill in establishing (4.14) in a given scenario is to find a suitable way to bound α and choosing an appropriate V such that (4.14) can be established. For the Random Walk Metropolis the kernel choice is such that $Q(x,dy) = q(|x-y|)dy$, meaning $\alpha(x,y) = 1 \wedge \pi(y)/\pi(x)$. Since the acceptance rate is just the ratio of target densities, it lends itself quite nicely to a simple bound. If we assume $\pi(x)$ is log-concave in the tails, then $\pi(y)/\pi(x) \leq \exp(-a(|y|-|x|))$ for large enough x. With this, a sensible choice of Lyapunov function would seem to be $V(x) = e^{s|x|}$, for some $0 < s < a$. We first consider the positive tail, i.e. the case $x \to \infty$. In this instance we can re-write the integral in (4.14) as

$$\begin{aligned}\int_{-\infty}^{0} [e^{s(|y|-x)} - 1]\alpha(x,y)Q(x,dy) + \int_{0}^{x} [e^{s(y-x)} - 1]\alpha(x,y)Q(x,dy) \\ + \int_{x}^{2x} [e^{s(y-x)} - 1]\alpha(x,y)Q(x,dy) + \int_{2x}^{\infty} [e^{s(y-x)} - 1]\alpha(x,y)Q(x,dy).\end{aligned}$$

The first and last terms can be made arbitrarily small by taking x large enough. In the first case this is because

$$\begin{aligned}\int_{-\infty}^{0} [e^{s(|y|-x)} - 1]\alpha(x,y)Q(x,dy) \leq \int_{-x}^{0} [e^{s(|y|-x)} - 1]Q(x,dy) \\ + \int_{-\infty}^{-x} [e^{s(|y|-x)} - 1]e^{-a(|y|-x)}Q(x,dy),\end{aligned}$$

where we have used the fact that $\alpha(x,y) \leq e^{-a(|y|-|x|)}$ for $|y| \geq |x|$. The first integral on the right-hand side is strictly negative for any x and the second is bounded above by $Q(x,(-\infty,-x))$, which will clearly become negligibly small as x grows. For the last term, we can again use the log-concave

restriction to bound the integral with

$$\int_{2x}^{\infty}[e^{(s-a)(y-x)}-e^{-a(y-x)}]Q(x,dy)\leq e^{(s-a)2x}Q(x,(2x,\infty))\to 0.$$

This leaves the middle two terms. We can combine these by writing $y=x+Z$, for $Z\sim\mu(\cdot)$, meaning $\mu(\cdot)$ denotes the zero mean proposal 'increment' distribution. Typically $\mu(\cdot)$ might be a zero mean Gaussian, if $Q(x,\cdot)=N(x,h\sigma^2)$. We can then bound the middle two integrals with

$$\int_{0}^{x}[e^{-sz}-1+e^{(s-a)z}+e^{-az}]\mu(dz)=-\int_{0}^{x}(1-e^{(s-a)z})(1-e^{-sz})\mu(dz), \tag{4.17}$$

which is strictly negative. Since for large x the entire integral will be comprised of terms which can be made arbitrarily small and terms which are strictly negative, this establishes (4.14) as $x\to\infty$, and the log-concave restriction means an equivalent argument holds as $x\to-\infty$.

4.2.2 Practical examples

In this section we give simple examples of the behaviour of some Metropolis–Hastings methods, in one dimension.

Independence sampler

As mentioned in Subsection 4.1.1, the Independence sampler suffers from the same problems in high dimensions as some other non-Markovian sampling methods. However, in one dimension it can perform very well in the case where the proposal distribution is a close approximation to the target, and provided it has heavier tails. Indeed, if the density $q(y)$ is a standard Gaussian and the same is true for $\pi(x)$, then the resulting Markov chain will actually consist of independent samples from $\pi(\cdot)$.

Figure 4.1, however, shows the case of a Gaussian proposal exploring a Cauchy target, with $\pi(x)\propto 1/(1+x^2)$. The left-hand plot shows that when the algorithm is initialised in the tails of the distribution it tends to get stuck there for large periods. This is because $q(y)/q(x)$ will be much larger than $\pi(y)/\pi(x)$ for inwards moves, so although most proposed moves will be inwards (provide the modes of $q(x)$ and $\pi(x)$ are close to one another) the acceptance rate $\alpha(x,y)=1\wedge\pi(y)/\pi(x)\times q(x)/q(y)$ will typically be very low for most of these. When the chain is initialised at the mode, the chain still fails to explore the distribution adequately, as evidenced by the right-hand plot, as $q(y)$ is such that values far away from zero are very unlikely to be proposed, whereas they still have a reasonable

chance of occurring under $\pi(\cdot)$. Crucially, this behaviour is extremely difficult to diagnose using convergence diagnostics like the graphs shown here without prior knowledge of the size of the typical set. Since we know the algorithm is not geometrically ergodic for this example, however, we have some understanding that the chain is unlikely to be a good representation of $\pi(\cdot)$.

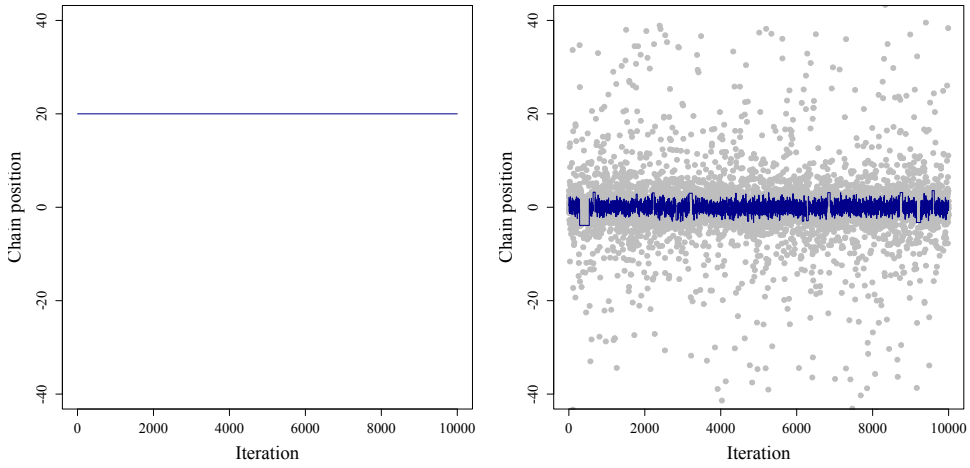

Figure 4.1: An independence sampler exploring a Cauchy target $\pi(x) \propto 1/(1+x^2)$ with a Gaussian proposal. The left-hand plot shows that when the chains is started in the tails it is likely to get stuck for long periods there. The right-hand plot shows the path of the chain (blue line) and independent samples from $\pi(\cdot)$ (grey dots), highlighting that the chain fails to adequately explore the typical set.

Random Walk Metropolis

Figure 4.2 shows behaviour of the Random Walk Metropolis with a Gaussian proposal exploring a Gaussian target, with $\pi(x) \propto e^{-x^2/2}$. The left-hand plot shows that convergence to the typical set is quite fast when the algorithm is initialised in the tails, owing to the acceptance rate $\alpha(x,y)$ ensuring that inwards moves are accepted and outwards moves mostly rejected. The middle plot shows that a chain initiated from within the typical set quickly explores the entirety of it. This is further emphasized by the histogram of samples from the chain in the right-hand plot, which closely resembles the Gaussian density which is plotted over the top of it.

By contrast, Figure 4.3 displays the behaviour of the same algorithm exploring the much heavier-

tailed target $\pi(x) \propto 1/(1+|x|^{1.1})$. The left-hand plot shows that when started in that tails, the chain tends to 'random walk', and fails to head in the direction of the mode for some time. This is because in this instance $\alpha(x,y) \approx 1$ for typical proposals in the tails, whether y is larger or smaller than x. The middle plot and histogram show that even when initialised at the mode, the method still fails to explore the distribution adequately, resulting in a biased histogram representation of the target density. Again, such behaviour is often difficult to diagnose without knowing the form that $\pi(x)$ should take.

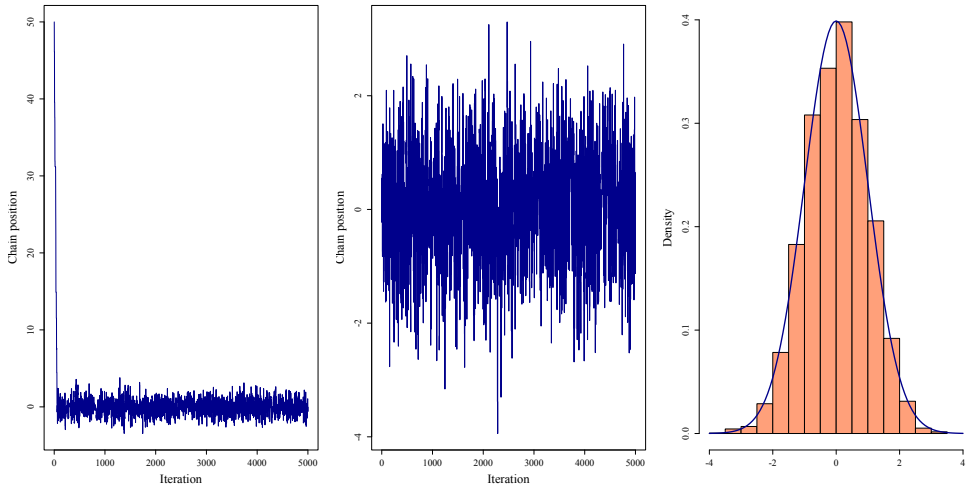

Figure 4.2: A Random Walk Metropolis exploring a Gaussian target $\pi(x) \propto e^{-x^2/2}$. When started in the tails (left-hand plot) the method quickly reaches the centre of the space, and from there it explores the distribution effectively (middle plot), as evidenced by the histogram which closely matches the overlaid Gaussian density (right-hand plot).

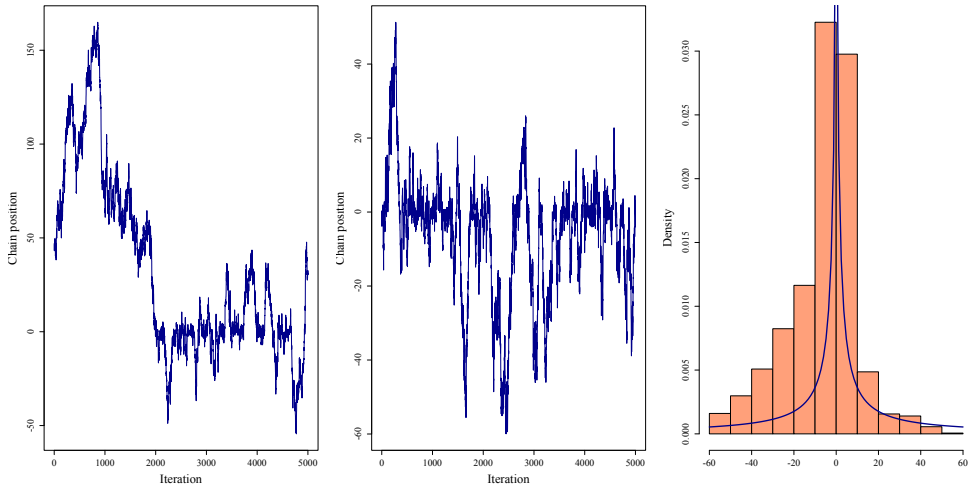

Figure 4.3: A Random Walk Metropolis exploring the target $\pi(x) \propto 1/(1+|x|^{1.1})$. The left-hand plot shows that when the chain is started in the tails it tends to 'random walk', and hence take a long time to reach the centre of the space. Once there the middle plot shows that it still fails to explore the distribution adequately, as evidenced by the skewed histogram (right-hand plot).

Metropolis-adjusted Langevin algorithm

On a form of $\pi(x)$ for which the tails are heavier than that of a Gaussian, MALA tends to perform better than the Random Walk Metropolis in one dimension (and often significantly better in more than one, owing to the better scaling discussed in Subsection 4.1.3). However, the algorithm still fails to be geometrically ergodic if $\pi(x)$ is not log-concave in the tails, as the gradient decays to zero here, meaning the algorithm still behaves as in the left-hand plot of Figure 4.3.

Figure 4.4 shows a different problem, which has also been discussed, that when $\pi(x)$ has lighter than Gaussian tails the gradients tend to 'explode' in the tails, meaning the algorithm can spend large periods there. When the current point in the chain is large, the proposal mass is very far from the mass under $\pi(\cdot)$, and so the majority of candidate moves will be rejected. In the example, when $x = 1$ then for a step-size $h = 1$ the proposal mean is $x - hx^3/2 = 1 - 1/2 = 1/2$. However when $x = 3$, as in the right-hand plot, then $x - hx^3/2 = 3 - 3^3/2 = -10.5$, which is far from zero.

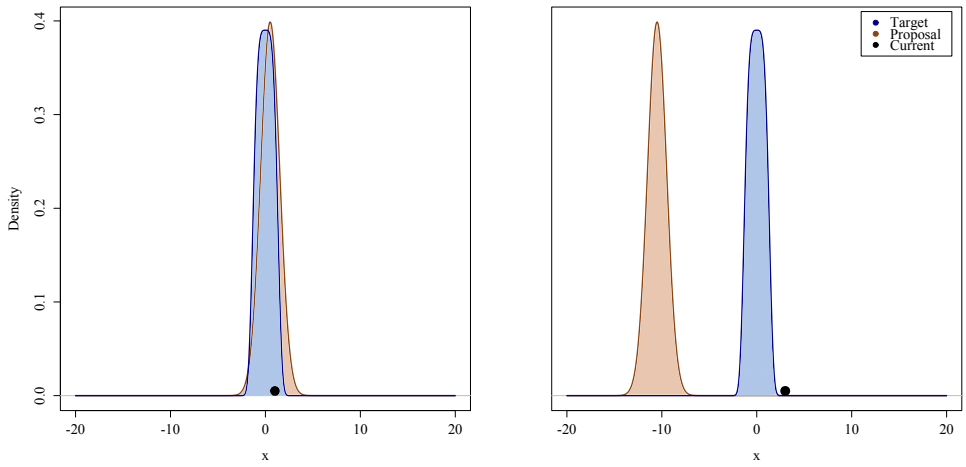

Figure 4.4: Metropolis-adjusted Langevin algorithm on a light-tailed target, $\pi(x) \propto e^{-x^4/4}$. When the current point x (black circle) is large, the proposal kernel (brown density) is a Gaussian centred at $x - hx^3/2$, which is very far from the typical set of the target density (blue), meaning most proposals will be rejected and the chain spends large periods in the tails.

Hamiltonian Monte Carlo

Empirical evidence shows that Hamiltonian Monte Carlo behaves in a similar way to MALA. This is intuitive: if the gradient becomes negligible in the tails then essentially the sampler will devolve into a random walk. Similarly if the target has very light tails then the gradient will 'explode' and the majority of proposals will be even further into the tails (and hence are likely to be rejected). Since the proposal kernel is much more complicated here, however, there is currently no theory to support this intuition. The reader is referred here to Chapter 7, where we provide some results in this direction, and explore the issues further.

4.3 Geometry in Markov chain Monte Carlo

This section is mostly taken from the author's published work [71]. Ideas from differential geometry have been successfully applied to statistics from as early as [53], offering new insight into common problems (e.g., [26, 75]). A survey is given in [5]. In this section, we suggest why some ideas from differential geometry may be beneficial for sampling methods based on Markov chains.

4.3.1 Manifolds and Markov chains

We often make assumptions in MCMC about the properties of the space, $\mathbf{X}$, in which our Markov chains evolve. Often $\mathbf{X} = \mathbb{R}^n$ or a simple re-parametrisation would make it so. However, here, $\mathbb{R}^n = \{(a_1, ..., a_n) : a_i \in (-\infty, \infty)\ \forall i\}$, the set of n-tuples of real numbers. The additional assumption that is often made is that $\mathbb{R}^n$ is *Euclidean*, an inner product space with the induced distance metric

$$d(x,y) = \sqrt{\sum_i (x_i - y_i)^2}. \tag{4.18}$$

For sampling methods based on Markov chains that explore the space locally, it may be advantageous to instead impose a different metric structure on the space, $\mathbf{X}$, so that some points are drawn closer together and others pushed further apart. Intuitively, one can picture distances in the space being defined such that if the current position in the chain is far from an area of $\mathbf{X}$ which is 'likely to occur' under $\pi(\cdot)$, then the distance to such a typical set could be reduced. Similarly, once this region is reached, the space could be 'stretched' or 'warped', so that it is explored as efficiently as possible.

While the idea is attractive, it is far from a constructive definition. We only have the pre-requisite that $(\mathbf{X}, d)$ must be a metric space. However, as many of the algorithms we have introduced use

gradient information, we will require $(\mathbf{X}, d)$ to be a space on which we can do differential calculus. Riemannian manifolds are an appropriate choice, therefore, as the rules of differentiation are well understand for functions defined on them (see Chapters 2 and 3 of [13]), while we are still free to define a more local notion of distance than Euclidean. In this section, we write $\mathbb{R}^n$ to denote the Euclidean vector space.

4.3.2 Geometry preliminaries

We do not provide a full overview of Riemannian geometry here (see [13, 64, 32]). We simply note that for our purposes, we can consider an n-dimensional Riemannian manifold (henceforth manifold) to be an n-dimensional metric space, in which distances are defined in a specific way. We also only consider manifolds for which a global coordinate chart exists, meaning that a mapping $r : \mathbb{R}^n \to M$ exists which is both differentiable and invertible and for which the inverse is also differentiable (a diffeomorphism). Although this restricts the class of manifolds available (the sphere, for example, is not in this class), it is again suitable for our needs and avoids the practical challenges of switching between coordinate patches. The connection with $\mathbb{R}^n$ defined through r is crucial for making sense of differentiability in M. We say a function $f : M \to \mathbb{R}$ is "differentiable" if $(f \circ r) : \mathbb{R}^n \to \mathbb{R}$ is (see Chapter 3 of [13]).

As has been stated, Equation (4.18) can be induced via a Euclidean inner product, which we denote $\langle \cdot, \cdot \rangle$. However, it will aid intuition to think of distances in $\mathbb{R}^n$ via curves

$$\gamma : [0,1] \to \mathbb{R}^n. \tag{4.19}$$

We could think of the distance between two points in $x, y \in \mathbb{R}^n$ as the minimum length among all curves that pass through x and y. If $\gamma(0) = x$ and $\gamma(1) = y$, the length is defined as

$$L(\gamma) = \int_0^1 \sqrt{\langle \gamma'(t), \gamma'(t) \rangle} dt, \tag{4.20}$$

where $\gamma' := d\gamma/dt$, giving the metric

$$d(x,y) = \inf \{ L(\gamma) : \gamma(0) = x, \gamma(1) = y \}. \tag{4.21}$$

In $\mathbb{R}^n$, the curve with a minimum length will be a straight line, so that Equation (4.21) agrees with Equation (4.18). More generally, we call a solution to Equation (4.21) a geodesic [13].

In a vector space, metric properties can always be induced through an inner product (which also gives a notion of orthogonality). Such a space can be thought of as "flat", since for any two points, y

and z, the straight line $ay+(1-a)z,\ a\in[0,1]$ is also contained in the space. In general, manifolds do not have vector space structure globally, but do so at the infinitesimal level. As such, we can think of them as "curved". We cannot always define an inner product, but we can still define distances through (4.21). We define a curve on a manifold, M, as $\gamma_M : [0,1] \to M$. At each point $\gamma_M(t) = p \in M$, the velocity vector, $\gamma'_M(t)$, lies in an n-dimensional vector space, which touches M at p. These are known as tangent spaces, denoted T_pM, which can be thought of as local linear approximations to M. We can define an inner product on each as $g_p : T_pM \to \mathbb{R}$, which allows us to define a generalisation of (4.20) as

$$L(\gamma_M) = \int_0^1 \sqrt{g_p(\gamma'_M(t), \gamma'_M(t))}dt. \tag{4.22}$$

and provides a means to define a distance metric on the manifold as

$$d(x,y) = \inf\{L(\gamma_M) : \gamma_M(0) = x, \gamma_M(1) = y\}.$$

We emphasise the difference between this distance metric on M and g_p, which is called a Riemannian metric or metric tensor and which defines an inner product on T_pM.

Embeddings and local coordinates

So far we have introduced manifolds as abstract objects. In fact, they can also be considered as objects that are embedded in some higher-dimensional Euclidean space. A simple example is any two-dimensional surface, such as the unit sphere, lying in $\mathbb{R}^3$. If a manifold is embedded in this way, then metric properties can be induced from the ambient Euclidean space.

We seek to make these ideas more concrete through an example, the graph of a function, $f(x_1,x_2)$, of two variables, x_1 and x_2. The resulting map, r, is

$$r : \mathbb{R}^2 \to M \tag{4.23}$$

$$r(x_1,x_2) = (x_1,x_2,f(x_1,x_2)). \tag{4.24}$$

We can see that M is embedded in $\mathbb{R}^3$, but that any point can be identified using only two coordinates, x_1 and x_2. In this case, each T_pM is a plane, and therefore, a two-dimensional subspace of $\mathbb{R}^3$, so: (i) it inherits the Euclidean inner product, $\langle\cdot,\cdot\rangle$; and (ii) any vector, $v \in T_pM$, can be expressed as a linear combination of any two linearly independent basis vectors (a canonical choice is the partial derivatives $\partial r/\partial x_1 =: r_1$ and r_2, evaluated at $x = r^{-1}(p) \in \mathbb{R}^2$). The resulting inner product, $g_p(v,w)$,

between two vectors, $v, w \in T_pM$, can be induced from the Euclidean inner product as

$$\begin{aligned}\langle v, w\rangle &= \langle v_1 r_1(x) + v_2 r_2(x), w_1 r_1(x) + w_2 r_2(x)\rangle, \\ &= v_1 w_1 \langle r_1(x), r_1(x)\rangle + v_1 w_2 \langle r_1(x), r_2(x)\rangle + v_2 w_1 \langle r_2(x), r_1(x)\rangle + v_2 w_2 \langle r_2(x), r_2(x)\rangle, \\ &= v^T G(x) w,\end{aligned}$$

where

$$G(x) = \begin{pmatrix} \langle r_1(x), r_1(x)\rangle & \langle r_1(x), r_2(x)\rangle \\ \langle r_1(x), r_2(x)\rangle & \langle r_2(x), r_2(x)\rangle \end{pmatrix} \tag{4.25}$$

and we use v_i, w_i to denote the components of v and w. To write (4.20) using this notation, we define the curve, $x(t) \in \mathbb{R}^2$, corresponding to $\gamma_M(t) \in M$ as $x = (r^{-1} \circ \gamma_M) : [0,1] \to \mathbb{R}^2$. Equation (4.20) can then be written

$$L(\gamma_M) = \int_0^1 \sqrt{x'(t)^T G(x(t)) x'(t)}\, dt, \tag{4.26}$$

which can be used in (4.21) as before.

The key point is that, although we have started with an object embedded in $\mathbb{R}^3$, we can compute the Riemannian metric, $g_p(v,w)$ (and, hence, distances in M), using only the two-dimensional "local" coordinates (x_1, x_2). We also need not have explicit knowledge of the mapping, r, only the components of the positive definite matrix, $G(x)$. The Nash embedding theorem [83] in essence enables us to define manifolds by the reverse process: simply choose the matrix, $G(x)$, so that we define a metric space with suitable distance properties, and some object embedded in some higher-dimensional Euclidean space will exist for which these metric properties can be induced as above. Therefore, to define our new space, we simply choose an appropriate matrix-valued map, $G(x)$ (we discuss this choice in Section 4.3.4). If $G(x)$ does not depend on x, then M has a vector space structure and can be thought of as "flat". Trivially, $G(x) = I$ gives Euclidean n-space.[1]

We can also define volumes on a Riemannian manifold in local coordinates. Following standard coordinate transformation rules, we can see that for the above example, the area element, dx, in $\mathbb{R}^2$ will change according to a Jacobian $J = |(Dr)^T (Dr)|^{1/2}$, where $Dr = \partial(p_1, p_2, p_3)/\partial(x_1, x_2)$. This reduces to $J = |G(x)|^{1/2}$, which is also the case for more general manifolds (see page 212 of [13]). We therefore define the Riemannian volume measure on a manifold, M, in local coordinates as

$$\mathrm{Vol}_M(dx) = |G(x)|^{\frac{1}{2}} dx. \tag{4.27}$$

[1] Note that the Euclidean space in which the n dimensional manifold can be embedded may not have $n+1$ dimensions. The Möbius strip and Klein bottles are two examples where $n+2$ are needed.

If $G(x) = I$, then this reduces to the Lebesgue measure.

4.3.3 Diffusions on manifolds

By a 'diffusion on a manifold' in local coordinates, we actually mean a diffusion defined on Euclidean space. For example, a realisation of Brownian motion on the surface, $S \subset \mathbb{R}^3$, defined in Figure 4.5 through $r(x_1,x_2) = (x_1,x_2,\sin(x_1)+1)$ will be a sample path, which is defined on S and 'looks locally' like Brownian motion in a suitably small neighbourhood of any point, $p \in S$. However, the pre-image of this sample path (through r^{-1}) will not be a realisation of a Brownian motion defined on $\mathbb{R}^2$, owing to the nonlinearity of the mapping. Therefore, to define *Brownian motion on S*, we define some diffusion $(X_t)_{t\geq 0}$ that takes values in $\mathbb{R}^2$, for which the process $(r(X_t))_{t\geq 0}$ 'looks locally' like a Brownian motion (and lies on S). See [73] for more intuition here.

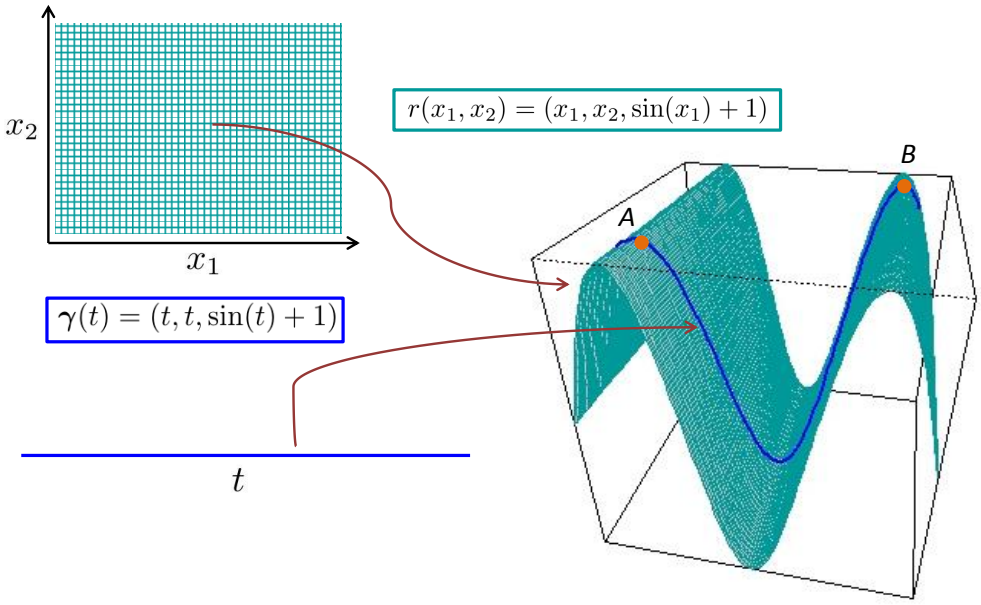

Figure 4.5: A two-dimensional manifold (surface) embedded in $\mathbb{R}^3$ through $r(x_1,x_2) = (x_1,x_2,\sin(x_1)+1)$, parametrised by the local coordinates, x_1 and x_2. The distance between points A and B is given by the length of the curve $\gamma(t) = (t,t,\sin(t)+1))$.

We can use the same intuition to define more general diffusions on manifolds, which we do in Chapter 5.

4.3.4 Choosing a metric

We now turn to the question of which metric structure to put on the manifold, or equivalently, how to choose $G(x)$. In this section, we sometimes switch notation slightly, denoting the target density, $\pi(x|y)$, as some of the discussion is directed towards Bayesian inference, where $\pi(\cdot)$ is the posterior distribution for some parameter, x, after observing some data, y. The goal is to find an appropriate choice of distance between points in the sample space of a given probability distribution.

A related (but distinct) problem is to define a distance between two probability distributions from the same parametric family, but with different parameters. This problem has been well-studied in *information geometry*, explored by Rao [94] and others (e.g. [1]) for many years. Although generic measures of distance between distributions (such as total variation) are often appropriate, based on information-theoretic principles, one can deduce that for a given parametric family, $\{p_x(y) : x \in \mathbf{X}\}$, it is in some sense natural to consider this 'space of distributions' to be a manifold, with the Fisher information as the Riemannian metric $G(x)$ (with the $\alpha = 0$ connection employed; see [1] for details).

Because of this, Girolami and Calderhead [43] proposed a variant of the Fisher metric for geometric Markov chain Monte Carlo, as

$$G_{ij}(x) = \mathbb{E}_{y|x}\left[-\frac{\partial^2}{\partial x_i \partial x_j}\log f(y|x)\right] - \frac{\partial^2}{\partial x_i \partial x_j}\log \pi_0(x), \tag{4.28}$$

where $\pi(x|y) \propto f(y|x)\pi_0(x)$ is the target density, f denotes the likelihood and π_0 the prior. The metric is tailored to Bayesian problems, so the Fisher information is combined with the negative Hessian of the log-prior. One can also view this metric as the expected negative Hessian of the log target with respect to Lebesgue measure, since this naturally reduces to (4.28).

The motivation for a Hessian-style metric can also be understood from studying MCMC proposals. For general pre-conditioning methods [108], the objective is to choose $G^{-1}(x)$ to match the covariance structure of $\pi(x|y)$ locally. If the target density were Gaussian with covariance matrix, Σ, then

$$-\frac{\partial^2}{\partial x_i \partial x_j}\log \pi(x|y) = \Sigma_{ij}. \tag{4.29}$$

In the non-Gaussian case, the negative Hessian is no longer constant, but we can imagine that it matches the correlation structure of $\pi(x|y)$ locally at least. Such ideas have been discussed in the geostatistics literature previously [22]. One problem with simply using (4.29) to define a metric

is that unless $\pi(x|y)$ is log-concave, the negative Hessian will not be globally positive-definite, although Petra *et al.* [93] conjecture that it may be appropriate for use in some realistic scenarios and suggest some computationally efficient approximation procedures [93].

Example 4.8. *Take* $\pi(x) \propto 1/(1+x^2)$*, and set* $G(x) = -\partial^2 \log \pi(x)/\partial x^2$*. Then,* $G^{-1}(x) = (1+x^2)^2/(2-2x^2)$*, which is negative if* $x^2 > 1$*, so unusable as a proposal variance.*

Girolami and Calderhead [43] use the Fisher metric in part to counteract this problem. Taking expectations over the data ensures that the likelihood contribution to $G(x)$ in (4.28) will be positive (semi-)definite globally (e.g. [90]); so, provided a log-concave prior is chosen, then (4.28) should be a suitable choice for $G(x)$. Indeed, Girolami and Calderhead [43] provide several examples in which geometric MCMC methods using this Fisher metric perform better than their 'non-geometric' counterparts.

Betancourt [8] also starts from the viewpoint that the Hessian (4.29) is an appropriate choice for $G(x)$ and defines a mapping from the set of $n \times n$ matrices to the set of positive-definite $n \times n$ matrices by taking a *smooth absolute value* of the eigenvalues of the Hessian. This is done in a way such that derivatives of $G(x)$ are still computable, inspiring the author to the name, *SoftAbs* metric. For a fixed value of x, the negative Hessian, $H(x)$, is first computed and, then, decomposed into $U^T D U$, where D is the diagonal matrix of eigenvalues. Each diagonal element of D is then altered by the mapping $t_\alpha : \mathbb{R} \to \mathbb{R}$, given by:

$$t_\alpha(\lambda_i) = \lambda_i \coth(\alpha \lambda_i), \tag{4.30}$$

where α is a tuning parameter (typically chosen to be as large as possible for which eigenvalues remain non-zero numerically). The mapping t_α acts as an absolute value function, but also uplifts eigenvalues which are close to zero to $\approx 1/\alpha$. It should be noted that while the Fisher metric is only defined for models in which a likelihood is present and for which the expectation is tractable, the SoftAbs metric can be found for any target distribution, $\pi(\cdot)$.

An important property of any Riemannian metric is how it transforms under coordinate change (e.g. [1]). The Fisher information metric commonly studied in information geometry is an example of a *coordinate invariant* choice for $G(x)$. If we consider two parametrisations for a statistical model given by x and $z = t(x)$, computing the Fisher information under x and then transforming

this matrix using the Jacobian for the mapping t, will give the same result as computing the Fisher information under z. It should be noted that because of either the prior contribution in (4.28) or the nonlinear transformations applied in other cases, none of the metrics we have reviewed here have this property, which means that we have no principled way of understanding how $G(x)$ will relate to $G(z)$. It is intuitive, however, that using information from all of $\pi(x)$, rather than only the likelihood contribution, $f(y|x)$, would seem sensible when trying to sample from $\pi(\cdot)$.

Chapter 5

Some new insights on Langevin diffusions

The Langevin diffusion $dX_t = \nabla \log \pi(X_t)dt + \sqrt{2}dB_t$ is a useful tool in Markov chain Monte Carlo as a relatively simple stochastic process which has a user-defined limiting distribution. But it seems natural to wonder whether there are other similar processes that do the same thing, which could therefore also be used as a basis for Markov chain sampling methods. By considering the Fokker–Planck equation (3.26), it can be seen that any diffusion $dX_t = b(X_t)dt + \sigma(X_t)dB_t$ with the drift b and volatility σ chosen such that

$$b_i(x) = \frac{1}{2\pi_u(x)} \sum_{j=l}^{n} \frac{\partial}{\partial x_j}[A_{ij}(x)\pi_u(x)],$$

will be π-invariant, where again $A(x) = \sigma(x)\sigma^T(x)$ is the 'squared' volatility and $\pi_u(x)$ denotes the unnormalised version of the density $\pi(x)$. So an infinite family of diffusions can be constructed for which $\pi(\cdot)$ is a stationary distribution, using only this equation. A sensible follow on question would therefore be whether there are more appropriate choices of diffusion to use as a basis for sampling than that with a constant volatility, on which the Metropolis-adjusted Langevin algorithm is based.

Many authors [108, 43] have in fact considered this natural extension to the Langevin diffusion (4.5) by allowing the volatility to vary with position. In [108], the authors suggest that such a diffusion can be constructed using the dynamics

$$dX_t = \frac{1}{2}G^{-1}(X_t)\nabla \log \pi(X_t)dt + \Omega(X_t)dt + \sqrt{G^{-1}(X_t)}dB_t, \tag{5.1}$$

with the ith component of the additional drift term given by

$$\Omega_i(X_t) = |G(X_t)|^{-1/2} \sum_{j=1}^{n} \frac{\partial}{\partial x_j}[G^{-1}_{ij}(X_t)|G(X_t)|^{\frac{1}{2}}]$$

The same equation is derived in [43], and the authors state that this is a generalisation of (4.5) to a Riemannian manifold with metric G. We establish here that in fact the diffusion (5.1) *does not* in general have invariant measure $\pi(\cdot)$.

The contributions of this chapter are three-fold. First, we highlight that the invariant distribution for (5.1) is not always $\pi(\cdot)$ and derive a simpler diffusion which does have the desired limiting probability measure using techniques from stochastic analysis. Secondly, we demonstrate how this simpler diffusion also naturally arises as a generalisation of (4.5) to a Riemannian manifold with metric G, using a more geometric approach. Third we discuss the convergence properties of the different diffusions, and how these relate to Markov chain sampling. The first part of the work is a collaboration with Tatiana Xifara, Christopher Sherlock, Simon Byrne and (along with the second) Mark Girolami. The next two sections are mainly paraphrased from [130] and [71].

5.1 Langevin diffusions with changing volatilities

If we begin with some positive definite, symmetric matrix-valued map $A : \mathbf{X} \to \mathbb{R}^{n\times n}$, we can construct the the diffusion

$$dX_t = b(X_t)dt + \sqrt{A(X_t)}dB_t,$$

where $\sqrt{A}$ denotes the matrix U such that $UU^T = A$. From here we can simply solve (3.26) to derive the correct form of b for which $\pi(\cdot)$ is invariant. Specifically we seek a drift $b(x)$ such that

$$b_i(x) = \frac{1}{2\pi(x)} \sum_j \frac{\partial}{\partial x_j}[A_{ij}(x)\pi(x)], \;\; i = 1, ..., n.$$

Solving gives

$$2b_i(x) = \sum_j A_{ij}(x)\frac{\partial}{\partial x_j}\log \pi(x) + \sum_j \frac{\partial A_{ij}}{\partial x_j}(x), \tag{5.2}$$

resulting in the diffusion

$$dX_t = \frac{1}{2}A(X_t)\nabla \log \pi(X_t)dt + \Lambda(X_t)dt + \sqrt{A(X_t)}dB_t, \tag{5.3}$$

where now the additional drift term has ith component

$$\Lambda_i(X_t) = \frac{1}{2}\sum_j \frac{\partial A_{ij}}{\partial x_j}(X_t).$$

We note that $\Lambda_i(x)$ is cheaper to compute than $\Omega_i(x)$ (and confirm this empirically in the next subsection). If we set $A(x) = G^{-1}(x)$ to match the notation of (5.1), then we have the following proposition.

Proposition 5.1. *If $G(x)$ is chosen such that for any combination of $1 \leq j,k,m \leq n$*

$$\frac{\partial}{\partial x_j}G_{km}(x) = \frac{\partial}{\partial x_k}G_{jm}(x) \tag{5.4}$$

for all x, then (5.1) and (5.3) represent the same diffusion.

Proof: Since the volatilities and the multipliers of $\nabla \log \pi$ in the drift are identical for the two diffusions, we need only show that $\Omega_i = \Lambda_i$ for all i. First we note that we can write

$$\begin{aligned}\Omega_i &= \sum_j \frac{\partial G^{-1}_{ij}}{\partial x_j} + \frac{1}{2}\sum_j G^{-1}_{ij}\frac{\partial}{\partial x_j}\log|G|, \\ &= -\sum_{j,k,m} G^{-1}_{ik}\frac{\partial G_{km}}{\partial x_j}G^{-1}_{mj} + \frac{1}{2}\sum_{j,k,m} G^{-1}_{ij}\frac{\partial G_{mk}}{\partial x_j}G^{-1}_{km},\end{aligned} \tag{5.5}$$

where we have used the general rule $\partial \log|G|/\partial x_j = \mathrm{tr}\left(G^{-1}\partial G/\partial x_j\right)$. From (5.4), the second term in (5.5) can be re-written

$$\frac{1}{2}\sum_{j,k,m} G^{-1}_{ij}\frac{\partial G_{jm}}{\partial x_k}G^{-1}_{km} = \frac{1}{2}\sum_{j,k,m} G^{-1}_{ik}\frac{\partial G_{km}}{\partial x_j}G^{-1}_{jm},$$

on relabelling $j \leftrightarrow k$. The result follows since $G^{-1}_{jm} = G^{-1}_{mj}$. ■

This property arises both when $n = 1$ and if G is the (continuous) Hessian matrix of some real-valued function, which goes some way towards explaining why the diffusion (5.1) was considered correct. In general, however (5.1) will not be π-invariant.

Theorem 5.2. *In general, the diffusion with dynamics governed by (5.1) will not have limiting distribution $\pi(\cdot)$.*

Proof: It suffices to construct a counter-example. For some positive-valued, differentiable function f, set

$$G(x) = \begin{pmatrix} f(x_2) & 0 \\ 0 & 1 \end{pmatrix}.$$

Then $\Lambda(x) = (0,0)^T$ and $\Omega(x) = (0, f'(x_2)/2f(x_2))^T$, and hence the diffusions (5.1) and (5.3) have different drift coefficients. Moreover, the diffusion (5.1) can be written in the same form as (5.3), and by matching drift terms it can be seen that its invariant density is actually proportional to $\pi(x)f(x_2)$.

■

5.1.1 Experiments

The following computer simulations were performed by Tatiana Xifara, not by this author, but are included here for completeness. The purpose was to compare two different Metropolis–Hastings schemes, one based on the diffusion (5.3), which we call 'PMALA' (position-dependent MALA), and another based on (5.1), which is known as 'MMALA' (manifold MALA). The comparison was performed across three of the scenarios considered in [43]: logistic regression on each of five different datasets, a stochastic volatility model, and a non-linear ODE model. As in [43] the Riemannian metric $G(x)$ was based on the expected Fisher information.

Initial tuning runs were performed to obtain the optimal scaling parameter h in terms of expected sample size (ESS) for each algorithm. The initialisation, burn-in, and length of each Markov chain was exactly as in [43], however here 100 (rather than 10) replications were performed for each chain.

Bayesian logistic regression and the non-linear ODE model are of most interest since in [43] MMALA was found to outperform Riemannian manifold Hamiltonian Monte Carlo for these scenarios. Detailed results are presented for the Bayesian logistic regression and non-linear ODE models. Results for the stochastic volatility model show the same underlying pattern. Where especially pertinent brief details on the models and the priors are given. For further details see [43].

Logistic regression

Here Bayesian logistic regression (e.g. [43]) was performed on five different datasets containing between 7 and 25 covariates. We choose a Gaussian prior for the parameter vector $\beta \sim N(0, \alpha I)$,

Dataset	*Method*	*ESS*	*CPU Time*	*min. ESS/s*
	PMALA	(685, 847, 986)	12.58	54.5
Australian Credit	MMALA	(696, 848, 943)	14.08	49.4
	PMALA	(605, 777, 917)	43.8	13.8
German Credit	MMALA	(605, 774, 921)	45.72	13.2
	PMALA	(659, 795, 923)	6.57	100.3
Heart	MMALA	(657, 773, 920)	8.07	81.4
	PMALA	(1235, 1415, 1572)	4.67	264.5
Pima Indian	MMALA	(1264, 1425, 1576)	5.59	226.1
	PMALA	(477, 591, 679)	3.32	143.7
Ripley	MMALA	(460, 590, 686)	3.94	116.7

Table 5.1: Results for two Metropolis-adjusted Langevin algorithms on a Bayesian logistic regression example. The mean (over the 100 replicates) is presented for the minimum, median and maximum ESSs (over the parameters). The CPU time and the mean minimum ESS per second are also given.

so that with a design matrix X and link function $s(\cdot)$ the metric tensor is given by $G(\beta) = X^T DX + \alpha^{-1} I$, where D is a diagonal matrix with elements $D_{i,i} = s(\beta^T X_{i,\cdot}^T)(1 - s(\beta^T X_{i,\cdot}^T))$. Under these assumptions the diffusions on which PMALA and MMALA are based have the same law and so the ESSs for these two algorithms should be the same up to Monte Carlo error.

For each Markov chain the ESS was computed for each parameter and the minimum, median and maximum of these was noted. Table 5.1 shows, for each algorithm and dataset, the means from 100 replicates. The CPU time and the mean (over replicates) minimum (over parameters) effective number of independent samples per second are also provided.

As expected, the ESSs for PMALA and MMALA are very similar. Since Λ is computationally less costly to calculate than Ω, PMALA is the quicker of the two algorithms and so obtains the largest ESS per second.

Non-linear differential equation model

A further model was considered based on the Fitzhigh-Nagumo differential equations in [95]: $\dot{W} = c\left(W - W^3/3 + R\right)$ and $\dot{R} = -(W - a + bT)/c$. The simulated dataset and our independent priors for the parameter vector (a, b, c) are the same as those used in [43]. To be consistent with the appendix of [43] and the associated Matlab code it was assumed that $\beta \sim \text{Exp}(1)$.

Method	*ESS*	*CPU Time*	*min. ESS/s*
PMALA	(1639.6, 669.3, 1406.4)	896.8	(1.83,0.75,1.57)
MMALA	(1274.4, 632.8, 1120.5)	923.0	(1.38,0.69,1.21)

Table 5.2: Results of the two MALA schemes for inference on the Fitzhugh-Nagumo model. For each parameter (a,b,c) and algorithm the mean (over the 100 replicates) ESS is presented, along with CPU time and mean minimum ESS per second.

The mean ESS for each parameter, along with its standard error are shown in Table 5.2, and it is clear that PMALA outperforms MMALA under this measure. CPU time and ESS/s are also provided in the table. PMALA is also the quickest algorithm, meaning its dominance is even clearer when CPU time is accounted for.

5.2 Langevin diffusions on manifolds

Our goal here is to define a diffusion on Euclidean space, which, when mapped onto a manifold through some diffeomorphism $r : \mathbb{R}^n \to M$, becomes the Langevin diffusion (4.5). Such a diffusion takes the form

$$dX_t = \frac{1}{2}\tilde{\nabla} \log \tilde{\pi}(X_t)dt + d\tilde{B}_t, \tag{5.6}$$

where those objects marked with a tilde must be defined appropriately.

We turn first to $(\tilde{B}_t)_{t\geq 0}$, which we use to denote Brownian motion on a manifold. Intuitively, we may think of a construction based on embedded manifolds, by setting $\tilde{B}_0 = p \in M$, and for each increment sampling some random vector in the tangent space T_pM, and then moving along the manifold in the prescribed direction for an infinitesimal period of time before re-sampling another velocity vector from the next tangent space [73]. In fact, we can define such a construction using Stratonovich calculus and show that the infinitesimal generator can be written using only local coordinates (e.g.

Section 5.5 of [111]). Here, we instead take the approach of generalising the generator directly from Euclidean space to the local coordinates of a manifold, arriving at the same result. We then deduce the stochastic differential equation describing $(\tilde{B}_t)_{t\geq 0}$ in Itô form using (3.25).

For a standard Brownian motion on $\mathbb{R}^n$, $\mathscr{A} = \Delta/2$, where Δ denotes the Laplace operator:

$$\Delta f = \sum_i \frac{\partial^2 f}{\partial x_i^2} = \text{div}(\nabla f). \tag{5.7}$$

Substituting $\mathscr{A} = \Delta/2$ into (3.25) trivially gives $b_i(x) = 0\ \forall i$, $A_{ij}(x) = \mathbb{1}_{\{i=j\}}$, as required. The Laplacian, $\Delta f(x)$, is the divergence of the gradient vector field of some function, $f \in C^2(\mathbb{R}^n)$, and its value at $x \in \mathbb{R}^n$ can be thought of as the average value of f in some neighbourhood of x [122].

To define a Brownian motion on any manifold, the gradient and divergence must be generalised. We provide a full derivation in Appendix C, which shows that the gradient operator on a manifold can be written in local coordinates as $\nabla_M = G^{-1}(x)\nabla$. Combining with the operator, div_M, we can define a generalisation of the Laplace operator, known as the Laplace–Beltrami operator (e.g. [51, 60]), as

$$\Delta_{LB} f = \text{div}_M(\nabla_M f) = |G(x)|^{-\frac{1}{2}} \sum_{i=1}^{n} \frac{\partial}{\partial x_i} \left(|G(x)|^{\frac{1}{2}} \sum_{j=1}^{n} G^{-1}_{ij}(x) \frac{\partial f}{\partial x_j} \right), \tag{5.8}$$

for some $f \in C_0^2(M)$.

The generator of a Brownian motion on M is $\Delta_{LB}/2$ [51]. Using (3.25), the resulting diffusion has dynamics given by

$$\begin{aligned} d\tilde{B}_t &= \Omega^*(X_t)dt + \sqrt{G^{-1}(X_t)}dB_t, \\ \Omega^*_i(X_t) &= \frac{1}{2}|G(X_t)|^{-\frac{1}{2}} \sum_{j=1}^{n} \frac{\partial}{\partial x_j} \left(|G(X_t)|^{\frac{1}{2}} G^{-1}_{ij}(X_t) \right). \end{aligned}$$

Those familiar with the Itô formula will not be surprised by the additional drift term, $\Omega^*(X_t)$. As Itô integrals do not follow the chain rule of ordinary calculus, non-linear mappings of martingales, such as $(B_t)_{t\geq 0}$, typically result in drift terms being added to the dynamics (e.g. Chapter 4 of [89]).

To define $\tilde{\nabla}$, we simply note that this is again the gradient operator on a general manifold, so $\tilde{\nabla} = G^{-1}(x)\nabla$. For the density, $\tilde{\pi}(x)$, we note that this density will now implicitly be defined with respect to the volume measure, $|G(x)|^{\frac{1}{2}}dx$, on the manifold. Therefore, to ensure the diffusion (5.6) has the correct invariant density with respect to the Lebesgue measure, we define

$$\tilde{\pi}(x) = \pi(x)|G(x)|^{-\frac{1}{2}}. \tag{5.9}$$

Putting these three elements together, Equation (5.6) becomes

$$dX_t = \frac{1}{2}G^{-1}(X_t)\nabla \log\left(\pi(X_t)|G(X_t)|^{-\frac{1}{2}}\right)dt + \Omega^*(X_t)dt + \sqrt{G^{-1}(X_t)}dB_t,$$

which, upon simplification, becomes

$$\begin{aligned} dX_t &= \frac{1}{2}G^{-1}(X_t)\nabla \log \pi(X_t)dt + \Lambda(X_t)dt + \sqrt{G^{-1}(X_t)}dB_t, \\ \Lambda_i(X_t) &= \frac{1}{2}\sum_j \frac{\partial}{\partial x_j}G^{-1}_{ij}(X_t). \end{aligned} \tag{5.10}$$

Intuitively, when a set is mapped onto the manifold, distances are changed by a factor, $\sqrt{G(x)}$. Therefore, to end up with the initial distances, they must first be changed by a factor of $\sqrt{G^{-1}(x)}$ before the mapping, which explains the volatility term in Equation (5.10).

The discrepency with this diffusion and (5.1) is that the latter is based on a 'Brownian motion on a manifold' with generator Δ_{LB} (without the $1/2$), and that it also has invariant density $\pi(x)$ with respect to the volume measure on the manifold, rather than Lebesgue measure.

5.3 Convergence properties

The main purpose of the previous sections was to correctly define a different class of Langevin diffusions. In this section we give some motivation for why basing Metropolis–Hastings methods on this class can be beneficial, in terms of the ergodic properties of the resulting samplers. We will discuss two methods. In the first, the proposal kernel takes the form

$$Q(x,\cdot) = N\left(x + \frac{h}{2}G^{-1}(x)\nabla \log \pi(x) + h\Lambda(x), hG^{-1}(x)\right). \tag{5.11}$$

We have previously referred to this as 'PMALA', standing for position-dependent Metropolis-adjusted Langevin algorithm. It is a straightforward Euler–Maruyama discretisation of the diffusion (5.3). In the second, the extra drift term $\Lambda(x)$ is ignored, leaving the proposal

$$Q(x,\cdot) = N\left(x + \frac{h}{2}G^{-1}(x)\nabla \log \pi(x), hG^{-1}(x)\right). \tag{5.12}$$

This is typically called the *simplified* manifold Metropolis-adjusted Langevin algorithm, or 'SMMALA' [43]. The additional drift term is ignored here simply to save on computing time.

Although there are a wealth of different choices available for the metric $G(x)$ (as discussed in both the previous and the next chapter), we focus on three specific cases here:

1. The negative Hessian, i.e.

$$G_1(x) = -\nabla^T \nabla \log \pi(x)$$

with suitable uplifting and absolute values taken of eigenvalues when necessary, in order to ensure that this matrix is positive-definite (to be used as a covariance)

2. The 'truncated' Metropolis-adjusted Langevin algorithm, in which

$$G_2(x) = ||\nabla \log \pi(x)||_\infty I_{n\times n},$$

where $I_{n\times n}$ denotes the $n \times n$ identity matrix and $\|x\|_\infty := \max_i |x_i|$ is the L_∞ norm. A version of this method was first introduced in [109]

3. A slightly less truncated version

$$G_3(x) = \|\nabla \log \pi(x)\|_\infty \text{diag}(x_i^{-1}),$$

where $\text{diag}(a_i)$ denotes an $n \times n$ diagonal matrix with ith diagonal element a_i.

Recall that the Langevin algorithm with drift $h\nabla \log \pi(x)/2$ fails to produce a geometrically ergodic chain either when $|\nabla \log \pi(x)| \to 0$ as $|x| \to \infty$ or $|\nabla \log \pi(x)|/|x| \to \infty$. In the former case proposals devolve into random walks, whilst in the latter they 'explode'. In this section we investigate whether any of these choices produce an algorithm which behaves more favourably in either of these scenarios.

The choice $G_1(x)$ is a generic form of a 'Hessian-style' metric, as first introduced and reviewed in Chapter 4. Recall that the motivation for the choice was to allow proposals to use local curvature information. There is a potentially $O(n^3)$ cost for inverting $G_1(x)$, which could feasibly be full rank.

The second and third choices $G_2(x)$ and $G_3(x)$ are simple attempts to control for faster-than-linear growth in $|\nabla \log \pi(x)|$. Dividing the drift by its maximum element will control this growth, making the resulting term $|G_2^{-1}(x)\nabla \log \pi(x)| = O(1)$, whereas $|G_3^{-1}(x)\nabla \log \pi(x) = O(|x|)$, i.e. a linear growth. In both cases the cost of computing $G^{-1}(x)$ is $O(n)$, as the matrices involved are diagonal. Rather than using new information, however, as in $G_1(x)$, derivative knowledge is simply recycled here.

We provide an intuitive discussion of the behaviour of proposals under each of the three metric choices, for two reference classes of targets in one dimension, and a specific distribution in two

dimensions. In one dimension the focus will be on stability of the proposals (5.11) and (5.12) for large $|x|$, and differences between them. In more than one dimension not only the *size* but also the *direction* of proposals (as characterised by the deterministic drift vector $b(x)h$ in the discretised diffusion) plays a role in the efficiency of samplers.

5.3.1 One dimension

In one dimension the first class of models we consider is a simplified version of the *exponential family* (discussed in greater detail in Chapter 7), with density

$$\pi(x) \propto \exp\left(-\beta^{-1}x^{\beta}\right), \quad x > 0,$$

for some $\beta > 0$. The case $\beta \geq 1$ implies log-concavity, while $\beta = 2$ implies Gaussian tails. The Metropolis-adjusted Langevin algorithm with fixed volatility produces a geometrically ergodic Markov chain in the case $1 \leq \beta \leq 2$. We note here that for the choice $G_1(x)$ a formal characterisation of geometric ergodicity for this class of targets and the proposal (5.12) is given in the comment [62], confirming that the resulting sampler produces a geometrically ergodic chain for any choice $\beta \neq 1$.[1] Here we provide some qualitative discussion to justify this result and compare with the proposal (5.11). This work was done independently of that in [62].

First note that $\nabla \log \pi(x) = -x^{\beta-1}$ here, which will shrink in the tails if $\beta < 1$ and grow at a faster than linear rate if $\beta > 2$, explaining the ergodicity results for MALA established in [109]. The necessary quantities for our purposes are given in Table 5.3 below. In the case $i = 1$ they are given for $\beta \neq 1$. Below, we comment on each metric choice in turn.

i	$G_i(x)$	$G_i^{-1}(x)\nabla \log \pi(x)$	$2\Lambda(x)$
1	$\lvert\beta-1\rvert x^{\beta-2}$	$-\lvert\beta-1\rvert^{-1}x$	$\frac{2-\beta}{\lvert\beta-1\rvert}x^{1-\beta}$
2	$x^{\beta-1}$	-1	$(1-\beta)x^{-\beta}$
3	$x^{\beta-2}$	$-x$	$(2-\beta)x^{1-\beta}$

Table 5.3: Gradient and curvature information of three different versions of the Metropolis-adjusted Langevin algorithm for the one-dimensional simplified *exponential family* class of models.

For $i = 1$, the first drift term will be linear and the second sublinear for any $\beta > 0$, meaning the

[1] In the case $\beta = 1$ then the Hessian is 0 so $G_1(x)$ is not defined.

diffusion will never be *stiff* and proposals will never 'explode' for large x, and similarly the drift will never become negligible in the tails. For $\beta > 1$ the second drift term will become negligible in the tails, meaning the two proposals (5.11) and (5.12) will become arbitrarily close to one another as $x \to \infty$. For $\beta < 1$ proposals will diverge in the tails, but the leading term will still be present in both. The second drift term will be positive for $\beta < 2$, effectively slowing down movement towards the centre of the space, and negative in the light-tailed case $\beta > 2$, speeding up the drift when there is very little mass in the tails. The volatility term will be $O(x^{1-\beta/2})$, which will always be sublinear in x for $\beta > 0$. For $\beta > 2$ this will effectively mean that proposals become deterministic in the tails. More discussion is given on volatility growth of proposals in Chapter 6.

For $i = 2$ the first drift term is always -1, meaning in the simplified proposal (5.12) proposals will effectively be a random walk with inwards drift. The second drift term will be less than the first for $x > 1$, and will always be negligible for large x, for any $\beta > 0$, making the two different proposals (5.11) and (5.12) arbitrarily similar here. The volatility will be $O(x^{(1-\beta)/2})$ which is again sublinear in x, and implies deterministic proposal behaviour in the tails for $\beta > 1$.

For the last case $i = 3$ the first drift term will again always be linear and the second sublinear for $\beta > 0$. For $\beta < 2$ the simplified proposal (5.12) will provide a stronger than optimal pull towards the centre of the space for large x, at a rate that increases as x does for $\beta < 1$ but decreases for $1 \leq \beta \leq 2$. The volatility here will be $O(x^{1-\beta/2})$, as in the case $i = 1$. Aside from constants, the terms for $i = 1$ and $i = 3$ are the same.

The second model we analyse is a simplified version of the *polynomial family* (discussed further in Chapter 6), with density

$$\pi(x) \propto x^{-p}, \quad x \geq 1,$$

for some $p > 1$ (note that $p = 2$ corresponds to Cauchy tails). In this case $\nabla \log \pi(x) = -p/x$ which becomes negligible in the tails, meaning the standard Metropolis-adjusted Langevin algorithm performs poorly here. The necessary terms required for the three metric choices are given in Table 5.4 below.

The first thing to note is that for $i = 1$ and $i = 3$ the resulting diffusions are identical for this class. So it seems that the necessary curvature information can be incorporated simply by intelligently recycling derivative information here. In these cases the first and second drift terms are linear, and

i	$G_i(x)$	$G_i^{-1}(x)\nabla \log \pi(x)$	$2\Lambda(x)$
1	px^{-2}	$-x$	$2x/p$
2	px^{-1}	-1	$1/p$
3	px^{-2}	$-x$	$2x/p$

Table 5.4: Gradient and curvature information of three different versions of the Metropolis-adjusted Langevin algorithm for the one-dimensional simplified *polynomial family* class of models.

the resulting diffusion will have dynamics

$$dX_t = (2/p-1)X_t dt + \sqrt{2X_t^2/p}dB_t.$$

For $p = 2$ (Cauchy tails) this will result in a diffusion without drift, and with volatility $\propto \pi(x)^{-1/2}$. In this case the diffusion is in fact equivalent to that arising from the class of *tempered* Langevin diffusions introduced in [108], and analysed in more detail in Chapter 6. For $1 < p < 2$ the simplified proposal will still give a strong drift towards the centre of the space, whereas the proposal (5.11) will actually drift away from the centre. For $p > 2$ the drift terms in both proposals will point towards the centre of the space, but the simplified scheme will pull more strongly towards the centre. The volatility term will always grow linearly in x. In fact, in this case the resulting diffusion will be exponentially ergodic here.

Proposition 5.3. *The diffusion with dynamics governed by the stochastic differential equation*

$$dX_t = (2/p-1)X_t dt + \sqrt{2x^2/p}dB_t$$

is exponentially ergodic to $\pi(x) \propto x^{-p}$ for $x > 1$, provided $p > 1+\varepsilon$ for some $\varepsilon > 0$.

Proof: Taking $V(x) = x^q$ for some $0 < q < 1$ chosen so that $\mathbb{E}_\pi[V(X)] < \infty$, then using (3.28) we have

$$\mathscr{A}V(x) = x^q \left[\left(\frac{2}{p} - 1 \right) q + \frac{q(q-1)}{p} \right].$$

We therefore need to show that the x^q multiplier is strictly negative. Simplifying gives the necessary and sufficient condition

$$p - 1 > q.$$

So provided $p > 1+\varepsilon$ then choosing $q < \varepsilon$ gives the result. ■

Discretisations of this diffusion without using Metropolis–Hastings corrections are likely to share these favourable properties (provided a small enough step-size is chosen), but convergence will typically be to an incorrect target distribution (e.g. [109]). It is unclear whether Metropolis–Hastings schemes based on this process would produce geometric converging chains, owing to the nonlinearities introduced through the acceptance probability. This issue is discussed in detail in the general case in [109] and [76]. We analyse Metropolis-corrected versions of a similar scheme in Chapter 6.

For the case $i = 2$ the first drift term will again be constant, as will the second. The combined drift will be $(1/p-1)$, which will always point towards the centre of the space for $p > 1$. Hence the simplified proposal will again propose moves which are closer to the centre of the space when in the tails, with the difference most severe for smaller p. The volatility here will be $O(x^{1/2})$, i.e. sublinear in x. The resulting proposal $y = x + h(1/p-1)/2 + x\sqrt{h/p}Z, Z \sim N(0,1)$, looks remarkably simple, however taking the same Lyapunov function here does not lead to a proof of exponential ergodicity.

5.3.2 Higher dimensions

In many examples of hierarchical models the resulting Langevin diffusion exhibits 'stiffness', meaning $|\nabla \log \pi(x)|/|x| \to \infty$ as $|x| \to \infty$ in at least one direction. A simple practical example is the Normal-Gamma model (e.g. [68]), in which the likelihood based on a sample with mean $\bar{x}$, variance s^2 and size b is given by

$$\pi(\tau, \mu) \propto \tau^{\frac{b}{2}} \exp\left(-\frac{\tau}{2}(bs^2 + b(\bar{x}-\mu)^2\right),$$

in which the leading order term is $O(\tau\mu^2)$. Fixing τ and letting μ grow indefinitely will result in exploding proposals here.

As an illustrative example we consider a model proposed in [110], with density

$$\pi(x) \propto \exp\left(-x_1^2 - x_2^2 - x_1^2 x_2^2\right).$$

As can be seen, the gradient vector is $\nabla \log \pi(x) = -2(x_1(1+x_2^2), x_2(1+x_1^2))$. Choosing the specific sequence $x_m = (m, 2)$, then the gradient vector becomes $\nabla \log \pi(x_m) = (-10m, -4(1+m^2))$, meaning $|\nabla \log \pi(x_m)|/|x_m| \to \infty$, and hence ordinary MALA proposals will explode in the tails, and the method will fail to produce a geometrically ergodic chain as a result.

For the sequence x_m we compare the behaviour of the three metric choices. Figure 5.2 gives some understanding of the behaviour of each. The first plot shows how $|b_i(x_m)|/|x_m|$ grows as $|x_m|$ does. While the standard Metropolis-adjusted Langevin algorithm exhibits faster than linear growth in drift, this is not the case for any of the other three metric choices, which all appear to be linear or sublinear. The second plot gives an idea of the extent to which each drift term points to the mode, by plotting $\langle b_i(x_m), -x_m \rangle$ against m. The first metric choice appears to produce a drift which asymptotically points towards the model $(0,0)$. The second is very close to being simply a normalized gradient, and so asymptotes towards $(0,-1)$. The third metric choice points somewhere in the middle of these two extremes. The last plot shows the ratio $|G_i^{-1}(x_m)\nabla \log \pi(x_m)|/|2\Lambda(x_m)|$, to understand how important the nonlinear term $\Lambda(x)$ is in determining each $b_i(x)$. It appears that in each case this ratio grows larger as m increases, particularly for the second metric choice.

Figure 5.3 also gives some qualitative intuition for how each method behaves. Below a contour plot are (unit) vector fields show the drift under each metric choice varies with position, with black showing G_1, red G_2 and green G_3 as in the other plots. It is clear that the first metric choice produces drift which points towards the global mode, the second choice does not when one coordinate is fixed and the other allowed to grow, and the third choice is a compromise between these.

To illustrate how magnitude and direction combine in this example the diffusions produced using each metric are shown in Figure 5.1. Here it is clear that the diffusion generated using Hessian information (G_1) reaches the centre of the space much more quickly than either of the other two choices (using a fixed time discretisation $h = 0.1$). Although informative, it should be noted that using the diffusions themselves does not directly translate to performance as a basis for a Metropolis–Hastings scheme, as it is likely that different optimal values of the time step h would be preferred for each method, meaning that using the same time step for each is not necessarily a fair comparison.

Although the Hessian-style metric seems favourable in many cases, and similar problems have been studied in the optimisation and information geometry literature (e.g. [1]), there are numerical challenges. While the second and third metric choices are clearly positive definite, the eigenvalues of $-\nabla^T \nabla \log \pi(x_m)$ are

$$\lambda_1(m) = 6 + m^2 - \sqrt{16 + 56m^2 + m^4}, \quad \lambda_2(m) = 6 + m^2 + \sqrt{16 + 56m^2 + m^4}.$$

Basic calculations show that $\lambda_1(m) \to -22$ as $m \to \infty$, meaning in practice the negative Hessian will not be positive definite for large m, and requires regularisation. This presents numerical challenges,

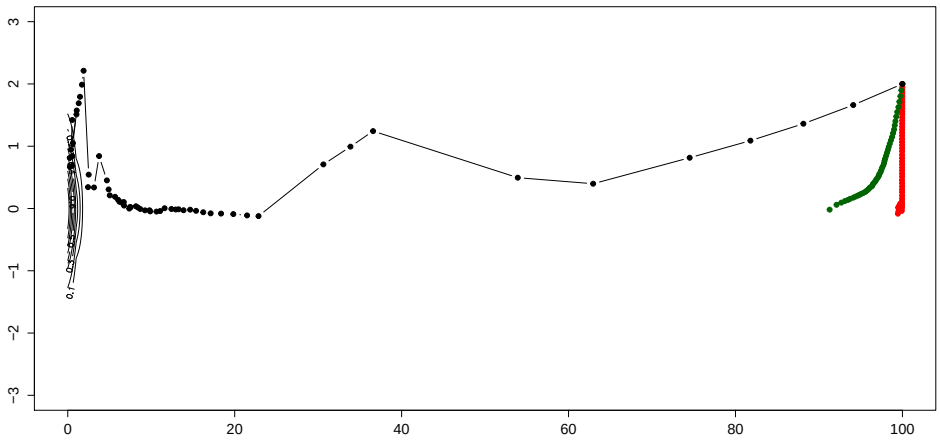

Figure 5.1: Discretisations of the Langevin diffusions resulting from the Hessian-style metric (black), truncating metric (red) and linearising metric (green).

particularly when $\lambda_1 \approx 0$, in which case $G_1^{-1}(x_m)$ can become extremely large if care is not taken to 'uplift' these eigenvalues by an appropriate amount.

5.4 Discussion & Extensions

The first part of this chapter is mainly concerned with correcting an error which has propagated through the literature. In particular, Theorem 5.2 gives rigorous justification for the use of the corrected diffusion. Empirical results further justify the corrected form, with the added benefit that it is computationally less expensive.

The second section is motivated by exploring the connections between differential geometry and stochastic analysis. It is not a new idea to relate changing the volatility of a diffusion to changing the space in which the diffusion exists, but there is value in better understanding this connection. For one, better understanding motivates new research questions, which can be tackled by a group of researchers (in this case geometers) who communicate in a different language. As both a more statistical and a more concrete example, in many cases Langevin diffusions arise as the limiting objects of Markov chain Monte Carlo methods (e.g. [104]), and recently such diffusion limits have been of the form described in this chapter, and the explicit geometric derivation has been acknowledged as a

useful pre-cursor to this result [7]. It is likely that future developments in the theory will also benefit from the geometric perspective.

Qualitative and intuitive exploration such as that given in the last section of this chapter is relevant to help translate the theory to practitioners, as well as to further understanding. For example, through the one dimensional exercise it becomes clear that for the class of models considered the omission of the nonlinear drift term $\Lambda(x)$ will not affect the ergodic properties of any sampler, and in some cases this term becomes negligible for large $|x|$. In fact, numerical methods exist to simulate the diffusion with $\Lambda(x)$ included without having to actually calculate it [14], and such schemes could be beneficial for Metropolis–Hastings sampling. It is also clear from this exercise that an appropriate metric choice $G(x)$ can change both the magnitude and direction of the drift vector $b(x)$, and hence the ergodic properties of both the diffusions and numerical schemes. In the examples shown, the magnitude properties of a metric which uses second derivative information about $\pi(x)$ can be recovered by judiciously recycling first derivative information, but in two dimensions, when direction also becomes important, the same cannot be said.

There are many interesting open questions on the topic of Langevin diffusions such as those discussed here. For the diffusions themselves, two obvious such questions are regarding speed of convergence to equilibrium and optimal metric choice to achieve this speed. Recent work [36] has studied spectral gaps of the generator $\mathscr{A}$ for related diffusions, which have unit volatility but additional drift components (making them nonreversible), and a similar formal analysis could give useful insights here on the speed of convergence to equilibrium for certain metric choices, as well as the related but distinct problem of minimising the asymptotic variance of estimators of functionals using the diffusion path (see [82] for more detail on the connections between these two problems). As an example, the recent paper [107] has established that in one dimension, if $\pi(x)$ has exponential tails and if $G_1(x) \geq G_2(x)$ for all $x \in \mathbf{X}$, then the diffusion with volatility $G_1^{-1/2}(x)$ will produce lower asymptotic variances for $L_2(\pi)$ functionals than that with volatility $G_2^{-1/2}(x)$. Through this we can see that relating volatility choice to estimator efficiency is possible and can lead to straightforward comparison criteria. Optimal metric choice is a more detailed and open question, though discussion in [36] and related papers suggests that progress can be made here too.

An alternative avenue to analysing the objects discussed in this chapter would be through hitting times to the centre of the space. A vast literature exists on hitting times for diffusions, with contributions from both the mathematical finance and partial differential equations literature, as well as

Probability theory, which could mean that such hitting times may be relatively straightforward to establish for many models of interest, giving explicit bounds on how long a diffusion takes to reach the mode of a distribution.

It is not always clear how analysis of diffusions translates to Metropolis–Hastings methods that use them as proposals. Indeed questions of 'speed' become more subtle here, as it may be that 'slower' diffusions can also be more accurately integrated numerically, meaning larger step-sizes can be taken whilst retaining a high chance of acceptance, cancelling out any weaknesses the process would possess in the continuous time setting. The question of optimal acceptance rate for any such algorithm is also more subtle here: is not clear whether the optimum should be 0.574 as in MALA, or indeed that an optimal rate independent of both x and $\pi(\cdot)$ can be established. Questions of algorithm efficiency as a function of n are also more involved.

Regarding ergodicity, the three metric choices discussed in the last section of the chapter should all produce geometrically ergodic Markov chains according to the findings of [109]. However, it would appear from the examples that the algorithms will converge to equilibrium at very different speeds in practice. Existing ergodicity results focus mainly on the magnitude of the drift vector $b(x)$, so some exploration of how the *direction* of this vector influences either the existence of a spectral gap or the size of this gap (or equivalently the geometric *rate*) could offer insight here. Some insights gained from the work [103] could be useful in this endeavour.

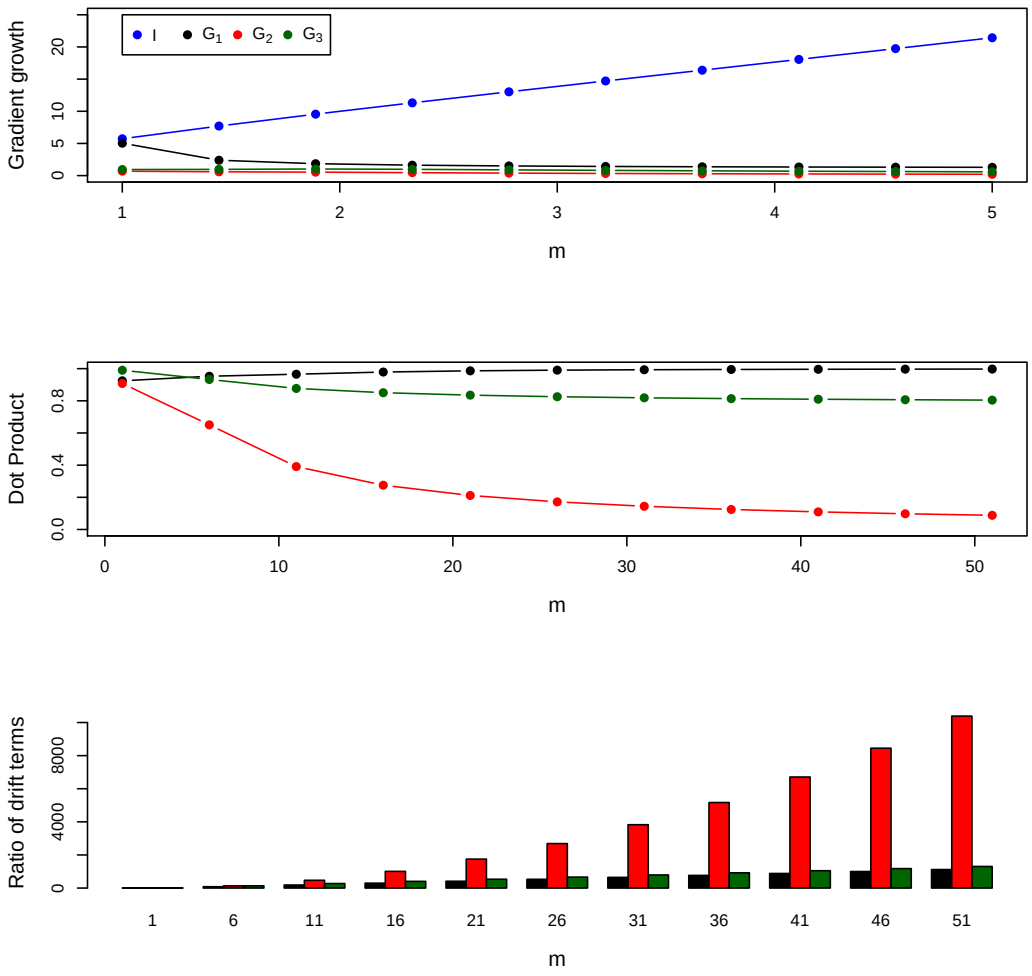

Figure 5.2: Plots showing behaviour of three Metropolis-adjusted Langevin algorithms for the target distribution $\pi(x) \propto \exp(-x_1^2 - x_2^2 - x_1^2 x_2^2)$. The first shows how the normalised drift terms $|b_i(x_m)|/|x_m|$ grow relative to $|x_m|$. The second compares the inner product $-\langle b_i(x_m), x_m \rangle$ with m. The third shows how the ratio of the first divided by the second drift terms changes with m.

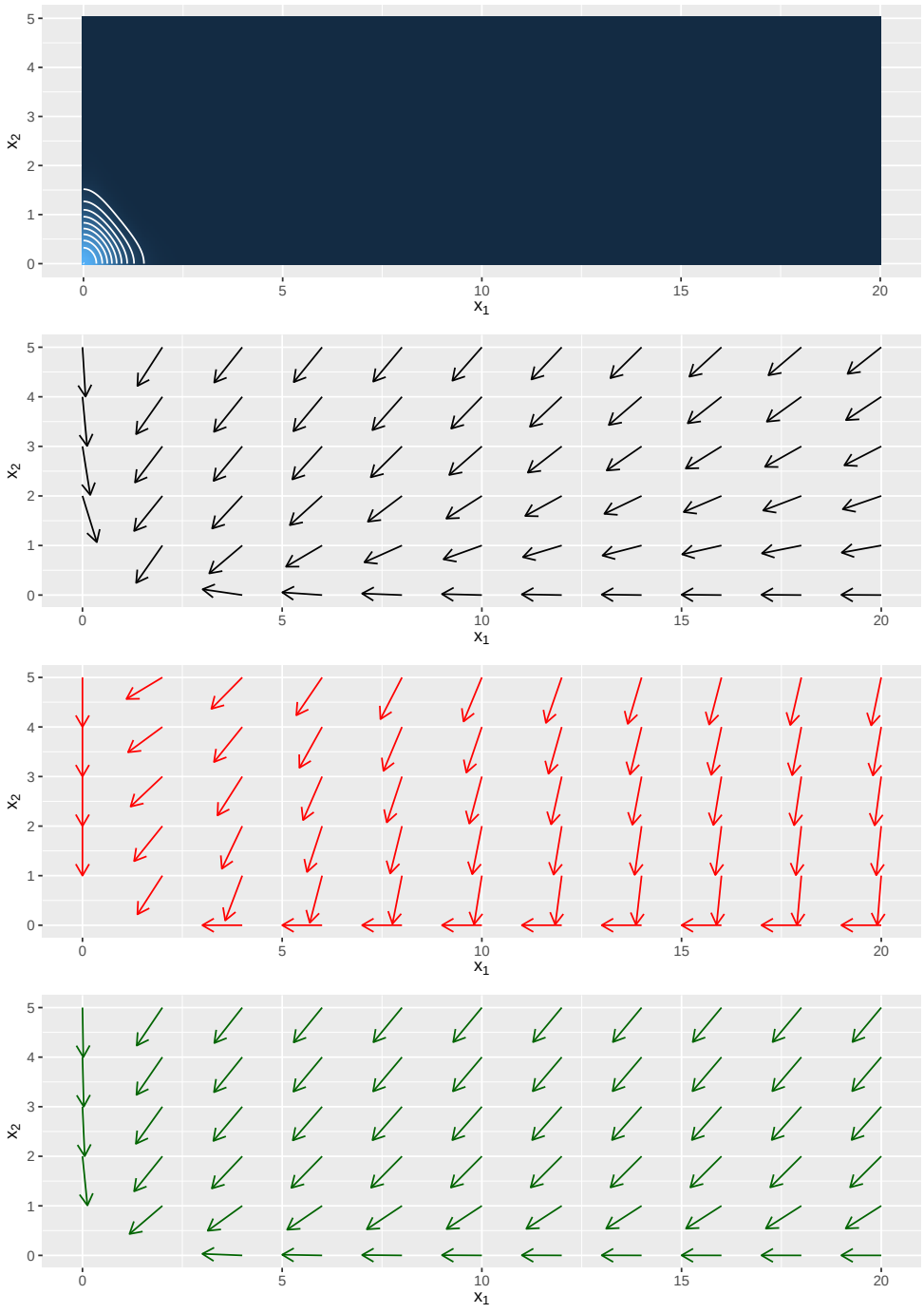

Figure 5.3: Vector fields showing the behaviour of each Metropolis-adjusted Langevin algorithm. The black lines represent the Hessian-style choice G_1, the red represents the truncated algorithm G_2 and green the linear growth variant G_3. The first graphic is a contour plot of the target density.

Chapter 6

Random walk Metropolis with position-dependent proposal covariance

This chapter is mainly taken from [70]. Although the research is solely the work of this author, it would have been much more difficult without regular discussions with Alexandros Beskos, as well as input from Krzysztof Łatuszyński and Gareth Roberts.

Recently, some MCMC methods have been proposed which generalise the Random Walk Metropolis described in Chapter 4, whereby proposals are still centred at the current point x and symmetric, but the variance changes with x [106, 108, 116, 3, 25]. The motivation is that the Markov chain can become more 'local', perhaps making larger jumps when out in the tails, or mimicking the local dependence structure of $\pi(\cdot)$ to propose more intelligent moves. Designing MCMC methods of this nature is particularly relevant for modern Bayesian inference problems, where posterior distributions are often high dimensional and exhibit nonlinear correlations [43]. We term this approach the *Position-Dependent Random Walk Metropolis* (PDRWM), although technically this is a misnomer, since proposals are no longer random walks.[1] Other choices of candidate distribution designed with

[1] The size of jump now depends on the current position in the chain.

distributions that exhibit nonlinear correlations were introduced in [43]. Although powerful, these require derivative information for $\log \pi(x)$, something which can be unavailable in modern inference problems (see e.g. Chapter 12 of [18]). We note that no such information is required for the PDRWM, as evidenced by the particular cases suggested in [106, 108, 116, 3, 25]. However, there are relations between the approaches, to the extent that understanding how the properties of the PDRWM differ from the standard RWM should also aid understanding of the methods introduced in [43].

In this work we consider the convergence rate of a Markov chain generated by the PDRWM to its limiting distribution. Our main interest lies in how much this generalisation can change these *ergodicity* properties compared to the standard RWM with fixed covariance. We focus on the case where the candidate distribution is Gaussian, and in one dimension we establish necessary and sufficient growth conditions on the proposal variance and tail behaviour of $\pi(x)$ for geometric ergodicity. Some of the results extend naturally to higher dimensions, but we also offer an illustrative example showing that some of the difficulties suffered by the RWM in dimensions two or greater can be alleviated when the proposal covariance is allowed to change with position.

General assumptions: As in previous chapters unless otherwise stated, we set $\mathbf{X} = \mathbb{R}^n$ here, so that objects such as Lebesgue densities and Gaussian measures are well understood. We also assume unless otherwise stated that the distribution of interest $\pi(\cdot)$ admits a Lebesgue density $\pi(x)$ which is bounded away from zero on compact sets.

6.1 Position-dependent Random Walk Metropolis

In the RWM, $Q(x,dy) = q(|y-x|)dy$, meaning the acceptance rate reduces to $\alpha(x,y) = 1 \wedge \pi(y)/\pi(x)$. A common choice is $Q(x,\cdot) = N(x, h\Sigma)$, with Σ chosen to mimic the global covariance structure of $\pi(\cdot)$ [121]. Various results exist concerning the optimal choice of h in a given setting (e.g. [104]). It is straightforward to see that Theorem 4.7 holds here, so that the tails of $\pi(x)$ must be uniformly exponential or lighter for geometric ergodicity. In one dimension this is in fact a sufficient condition [78], while for higher dimensions additional conditions are required [110]. We return to this case in Subsection 6.3.

For the PDRWM we introduce the matrix-valued map

$$G^{-1} : \mathbf{X} \to M^n_{PD}(\mathbb{R}),$$

where $M^n_{PD}(\mathbb{R})$ is defined as the space of $n \times n$ positive-definite matrices with real coefficients. Unless otherwise stated we assume that the eigenvalues of G^{-1} are bounded away from zero, to ensure that this matrix remains positive-definite for all $x \in \mathbf{X}$. So the transition kernel for the method is given by $Q(x,\cdot) = N(x, hG^{-1}(x))$, and the acceptance rate becomes

$$\alpha(x,y) = 1 \wedge \frac{\pi(y)|G(y)|^{\frac{1}{2}}}{\pi(x)|G(x)|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-y)^T[G(y)-G(x)](x-y)\right).$$

The intuition here is that proposals are more able to reflect the local dependence structure of $\pi(\cdot)$. In some cases this dependence may vary greatly in different parts of the state-space, making a global choice of Σ ineffective [116].

Readers familiar with differential geometry will recognise the volume element $|G(x)|^{1/2}dx$ and the linear approximations to the distance between x and y taken at each point through $G(x)$ and $G(y)$ if $\mathbf{X}$ is viewed as a Riemannian manifold with metric G.

The choice of $G(x)$ is an obvious question. In fact, specific variants of this method have appeared on many occasions in the literature, some of which we now summarise.

1. *Tempered Langevin diffusions* [108] $G^{-1}(x) = \pi^{-1}(x)I$. The authors highlight that the diffusion with dynamics $dX_t = \pi^{-\frac{1}{2}}(X_t)dB_t$ has invariant distribution $\pi(\cdot)$, motivating the choice. The method was shown to perform well for a bi-modal $\pi(x)$, as larger jumps are proposed in the low density region between the two modes.

2. *State-dependent Metropolis* [106] $G^{-1}(x) = a(1+|x|)^b$. Here the intuition is simply that $b > 0$ means larger jumps will be made in the tails. In one dimension the authors compare the expected squared jumping distance $\mathbb{E}[(X_{i+1} - X_i)^2]$ empirically for chains exploring a $N(0,1)$ target distribution, choosing b adaptively, and found $b \approx 1.6$ to be optimal.

3. *Regional adaptive Metropolis–Hastings* [106, 25]. $G^{-1}(x) = \sum_{i=1}^m \mathbb{1}_{x\in\mathbf{X}_i}\Sigma_i$. In this case the state-space is partitioned into $\mathbf{X}_1 \cup \ldots \cup \mathbf{X}_m$, and a different proposal covariance Σ_i is learned adaptively in each region $1 \leq i \leq m$. An extension which allows for some errors in choosing an appropriate partition is discussed in [25]

4. *Localised Random Walk Metropolis* [3]. $G^{-1}(x) = \sum_{k=1}^{m} \check{q}_\theta(k|x)\Sigma_k$. Here $\check{q}_\theta(k|x)$ are weights based on approximating $\pi(x)$ with some mixture of Normal/Student's t distributions, using the approach suggested in [2]. At each iteration of the algorithm a mixture component k is sampled from $\check{q}_\theta(\cdot|x)$, and the covariance Σ_k is used for the proposal $Q(x,dy)$.

5. *Kernel adaptive Metropolis–Hastings* [116]. $G^{-1}(x) = \gamma^2 I + \nu^2 M_x H M_x^T$, where $M_x = 2[\nabla_x k(z_1,x),...,\nabla_x k(z_n,x)]$ for some kernel function k and n past samples $\{z_1,...,z_n\}$, $H = I - 1/n\mathbb{1}_{n\times n}$ is a centering matrix, and γ, ν are tuning parameters. The approach is based around performing nonlinear principal components analysis on past samples from the chain to learn a local covariance. Illustrative examples for the case of a Gaussian kernel show that $M_x H M_x^T$ acts as a weighted empirical covariance of samples z, with larger weights given to the z_i which are closer to x [116].

The latter cases also motivate any choice of the form

$$G^{-1}(x) = \sum_{i=1}^{n} w(x,z_i)(z_i - x)^T(z_i - x)$$

for some past samples $\{z_1,...,z_n\}$ and weight function $w : \mathbf{X}\times\mathbf{X} \to [0,\infty)$ with $\sum_i w(x,z_i) = 1$ that decays as $|x - z_i|$ grows, which would also mimic the local curvature of $\pi(\cdot)$ (taking care to appropriately regularise and diminish adaptation so as to preserve ergodicity, as outlined in [3]). The logic of [43, 8] could also be applied, by choosing $G(x)$ as some regularised version of the negative Hessian of $\log\pi(x)$. However, if such derivative information were available it would seem more sensible to use a more sophisticated method than a martingale proposal (see e.g. [43]).

6.2 Geometric ergodicity in one dimension

Here the specific choice of $G(x)$ is left open, and we instead consider two different general scenarios as $|x| \to \infty$, i) $G^{-1}(x) \to \Sigma$, and ii) $G^{-1}(x) \to \infty$ at some rate. In theory there is also the possibility that $G^{-1}(x) \to 0$, though intuitively this would not seem to be a particularly sensible choice as chains would be extremely likely to spend a long time in the tails of a distribution, so we do not consider it.

Three scenarios are considered for the tail behaviour of $\pi(x)$. We refer to this density as *log-concave in the tails* if for some $x_0 > 0$ and $a > 0$

$$\pi(y)/\pi(x) \leq e^{-a(y-x)}, \quad \forall y \geq x \geq x_0, \tag{6.1}$$

and a similar condition holds in the negative tail. If (6.1) is not satisfied but there is some $\beta \in (0,1)$ such that the above condition can be replaced with $\pi(y)/\pi(x) \leq \exp\{-a(y^\beta - x^\beta)\}$, then we call the density subexponential (note this is not the standard definition). Finally, we call $\pi(x)$ 'polynomial-tailed' if $\pi(x) \propto |x|^{-p}$ for large $|x|$ and some $p \geq 1$. We also apply asymptotic growth conditions for $G^{-1}(x)$, and without loss of generality assume that these hold for any x larger than the same x_0 in absolute value.

We introduce some asymptotic notation in this section. For positive real-valued functions f and g, let $f(x) = \Theta(g(x))$ imply $f(x)/g(x) \to C > 0$ as $x \to \infty$, and $f(x) = \omega(g(x))$ imply $f(x)/g(x) \to \infty$. The more familiar big-O and little-o notation is also used. The main results of this section are summarised in Table 1 at the end of the section.

The first result emphasises a growing variance as a necessary requirement for geometric ergodicity in the heavy-tailed case.

Proposition 6.1. *If $G^{-1}(x) \leq \sigma^2$, then the PDRWM can produce a geometrically ergodic Markov chain* only *in the case where $\pi(x)$ is log-concave in the tails.*

Proof: In this case for any choice of $\varepsilon > 0$ there is a $\delta > 0$ such that $Q(x, B_\delta(x)) > 1 - \varepsilon$, so Theorem 4.7 can be applied. ■

Though the heavy-tailed case is a challenging scenario, the standard RWM with fixed covariance will produce a geometrically ergodic Markov chain if $\pi(x)$ is log-concave. Next we extend this result to the case of sub-quadratic variance growth in the tails.

Theorem 6.2. *If $G^{-1}(x) = o(|x|^2)$ and $\pi(x)$ is log-concave in the tails, then the PDRWM method produces a geometrically ergodic Markov chain from π-almost any starting point. If $\pi(x)$ is subexponential for some $\beta \in (0,1)$, then choosing $G^{-1}(x) = \Theta(|x|^\gamma)$ for some $2(1-\beta) < \gamma < 2$ gives the same result.*

The log-concave proof consists of partitioning $\mathbf{X}$ into five regions, and showing that as $|x| \to \infty$, (4.14) evaluated over each of these regions will either become arbitrarily small or remain strictly negative. We use the Lyapunov function $V(x) = e^{s|x|}$ for some $s > 0$. This choice allows results about

moment generating functions of truncated Gaussian distributions (see D) to be used, in conjunction with simple bounds on the cumulative distribution function from [24], to establish that (4.14) will become arbitrarily small for regions of $\mathbf{X}$ outside the 'typical set' $(x - cx^{\gamma/2}, x + cx^{\gamma/2})$. Theorem 3.2 from [78] shows that for the RWM with fixed covariance (4.14) evaluated over this region will be strictly negative. The essence of the argument is that for $y > x$ in the tails, $\alpha_R(x,y) \leq e^{-a(y-x)}$ by log-concavity, so as long as s is chosen to be less than a this decay will dominate any growth in $V(y)$ here. As for any inwards proposals $\alpha_R(x,y) = 1$ then it can be shown that (4.14) is strictly negative when evaluated over this region.

The crucial additional difficulty in the case of growing covariance is that the acceptance rate in this region (for suitably large x) is now

$$\alpha(x,y) = 1 \wedge \frac{\pi(y)}{\pi(x)} \exp\left(\frac{\gamma}{2}\log\left|\frac{x}{y}\right| - \frac{1}{2h}\left[\frac{(x-y)^2}{y^\gamma} - \frac{(x-y)^2}{x^\gamma}\right]\right)$$

The problematic term lies inside the square bracket: this will be negative for $y > x$, meaning a large positive component in $\alpha(x,y)$. To deal with this, we use a Taylor expansion of $y^{-\gamma}$ about x and some simplifications to show that provided $\gamma < 2$, for large enough x, *locally* (for y near x, where the choice of region plays a role) the acceptance rate will still satisfy

$$\alpha(x,y) = 1 \text{ for } y < x, \;\; \alpha(x,y) \leq e^{-a(y-x)+\delta_x}, \text{ for } y > x,$$

where δ_x can be made arbitrarily small. This allows us to use a similar argument to that in [78] to prove the result. Outside of this region the Gaussian tails of $Q(x,\cdot)$ take care of any less desirable behaviour of $\alpha(x,y)$. To extend this result to the subexponential case, we choose $V(x) = e^{s|x|^\beta}$, and Taylor expand $|y|^\beta$ in the typical set to get a suitable bound on $\alpha(x,y)$.

Note that this Theorem includes as a special case any instance in which $G^{-1}(x) \uparrow \sigma^2$ as $|x| \to \infty$. However, the case $G^{-1}(x) \to \sigma^2$ from any direction is actually more straightforward to show, by simply moving x far enough into the tails that $G^{-1}(x) \approx \sigma^2$ for all $y \in (x - cx^{\gamma/2}, x + cx^{\gamma/2})$. In this case the argument in [78] can be applied more straightforwardly.

Although we do not formally prove that the method will not produce a geometrically ergodic chain in the polynomial tailed case when $G^{-1}(x) = o(|x|^2)$, we show intuitively that this will be the case. Assuming that in the tails $\pi(x) \propto |x|^{-p}$ for some $p > 1$ then for large x

$$\alpha(x, x + cx^{\gamma/2}) = 1 \wedge \left(\frac{x}{x + cx^{\gamma/2}}\right)^{p+\gamma/2} \exp\left(-\frac{c^2 x^\gamma}{2h}\left[\frac{1}{(x+cx^{\gamma/2})^\gamma} - \frac{1}{x^\gamma}\right]\right).$$

The first expression on the right hand side converges to 1 as $x \to \infty$, which is akin to the case of fixed proposal covariance. The second term will be larger than one for $c > 0$ and less than one for $c < 0$. So the algorithm will exhibit the same 'random walk in the tails' behaviour which is often characteristic of the RWM in this scenario, and so the acceptance rate will fail to enforce a geometric drift back into the centre of the space.

In the case where $\gamma = 2$ this will not happen, as the terms in the above expression will be roughly constant with x. We examine this case next.

Theorem 6.3. *If $G^{-1}(x) = \Theta(|x|^2)$, then there is a $h_0 = h_0(G^{-1}) > 0$ such that for a step-size $h \in (0, h_0)$ the PDRWM method produces a geometrically ergodic Markov chain from π-almost any starting point, provided $\pi(x) \leq |x|^{-p}$ for all $|x| \geq L$, where $L < \infty$, for some $p > 1$.*

Here the intuition is that proposals in the tails will take the form $y = (1 + \xi\sqrt{h})x$, which if h is chosen to be small will be similar to $y = e^{\xi\sqrt{h}}x$. The latter scheme is sometimes called the *multiplicative* RWM, and is known to be geometrically ergodic in this scenario (e.g. [121]), as this equates to taking a log-transformation of x, which 'lightens' the tails of the target density to the point where it becomes log-concave.

In this case we take the Lyapunov function $V(x) = 1 \vee |x|^s$, with $s > 0$ chosen such that $\int V(y)\pi(dy) < \infty$. We again divide the integral of interest into regions, but in this case we show that each of these can be appropriately bounded simply as functions of the step-size h, i.e. independently of x. By examining each term, we show that for a small enough h the integral will be strictly negative.

The result is positive, but in this case is perhaps an example where the theory does not necessarily translate into an effective scheme in practice. If $\pi(x)$ has particularly heavy tails, for example, then it is likely that an extremely small value of h would be needed to ensure (3.18), meaning the geometric rate of convergence r would be close to one. Nonetheless, it is an example of how appropriate choice of $G^{-1}(x)$ can *favourably* change the ergodicity properties of a sampler.

The final result of this section provides a note of warning, that lack of care in choosing $G^{-1}(x)$ can have severe consequences for the method.

Theorem 6.4. *If $G^{-1}(x) = \omega(|x|^2)$, then the PDRWM method can never produce a geometrically*

ergodic Markov chain provided $\pi(y) \leq \pi(x)$ *for all* $|y| \geq |x| \geq L$*, for some* $L < \infty$*.*

The intuition for this result is straightforward when explained. In the tails, the average proposals will be of size $|x|^{\gamma/2}$, which will be much larger than $|x|$ if $\gamma > 2$, meaning most will send the chain even further into the tails in either direction (and hence will likely be rejected). To make this rigorous we show that (4.15) holds here, by considering the set of proposals $A_{x,\varepsilon} := \{y \in \mathbf{X} : \alpha(x,y) \geq \varepsilon\}$, and showing that $Q(x, A_{x,\varepsilon}) \to 0$ as $|x| \to \infty$, for any $\varepsilon > 0$. A specific example is illustrated in Figure 6.1.

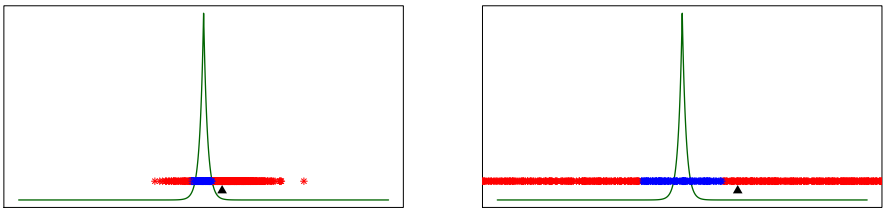

Figure 6.1: Example of Position-dependent Random Walk Metropolis behaviour with $\pi(x) \propto e^{-|x|}$, $G^{-1}(x) \propto |x|^4$. The black triangle denotes the current state, points highlighted in blue represent proposals with $\alpha(x,y) > 0.5$, with all others highlighted in red. For large $|x|$ the majority of proposals miss the centre of the space and are rejected.

The main results of this section are summarised in Table 6.1.

Variance	Polynomial Tails	Subexponential	Log-concave
$G^{-1}(x) = o(\|x\|^2)$	$\times$	$\checkmark^+$	$\checkmark$
$G^{-1}(x) = \Theta(\|x\|^2)$	$\checkmark^*$	$\checkmark^*$	$\checkmark^*$
$G^{-1}(x) = \omega(\|x\|^2)$	$\times$	$\times$	$\times$

Table 6.1: Summary of one dimensional ergodicity results for Position-dependent Random Walk Metropolis. Here $f(x) = \omega(g(x))$ means $f/g \to \infty$ as $x \to \infty$, $f(x) = \Theta(g(x))$ means $f/g \to C > 0$, $\checkmark$means geometrically ergodic, $\checkmark^+$ means geometrically ergodic provided $G^{-1}(x) \in \Theta(|x|^\gamma)$ for some $2 > \gamma > 2(1-\beta)$, and $\checkmark^*$ means geometrically ergodic provided h is suitably small.

6.3 Higher dimensions

Some results from the previous section naturally carry over to higher dimensions. The most straightforward is outlined below.

Proposition 6.5. *If each element of $G^{-1}(x)$ is bounded above (uniformly in x), then the PDRWM can* only *produce a geometrically ergodic Markov chain if the tails of $\pi(x)$ are uniformly exponential or lighter.*

Proof: As with Proposition 6.1, a straightforward application of Theorem 4.7 gives the result. ■

It is also intuitive that an analogue to Theorem 6.4 will exist here. Specifically, if any diagonal component of the covariance $G^{-1}(x)$ grows at a faster than quadratic rate with x, then the sampler is likely to run into the same difficulties in the tails. Similarly, when $G^{-1}(x) \to \Sigma$, it is straightforward to see that the sampler will inherit the geometric ergodicity properties of the RWM with fixed covariance, by a similar argument to that discussed for the proof of Theorem 6.2 in this case.

As mentioned earlier, in the case $G^{-1}(x) = \Sigma$, additional conditions on $\pi(x)$ are required for geometric ergodicity in more than one dimension, outlined in [110]. An example is also given in the paper of the simple two-dimensional density $\pi(x,y) \propto \exp(-x^2 - y^2 - x^2y^2)$, which fails to meet this criterion. The difficult models are those for which probability concentrates on a 'ridge' in the tails, which becomes ever narrower as $|x|$ increases. In this instance, proposals from the RWM are less and less likely to be accepted as $|x|$ grows. The problem is illustrated graphically in Figure 6.2. Such densities are often encountered as posterior distributions in hierarchical models, with another well-known example being the 'funnel', discussed in [85]. On the same figure there is some graphical evidence that if the proposal covariance is allowed to adjust then this problem can be alleviated somewhat.

To explore this more concretely, we design an extremely simple two dimensional density which exhibits the same features, which we call the 'rectangle' density

$$\square(x) \propto 3^{-\lfloor x_2 \rfloor} \mathbb{1}_R(x), \;\; R := \{y \in \mathbb{R}^2; y_2 \geq 1, |y_1| \leq 3^{1-\lfloor y_2 \rfloor}\},$$

where $\lfloor z \rfloor$ is the integer part of $z \in \mathbb{R}$. This is simply a distribution defined over a sequence of

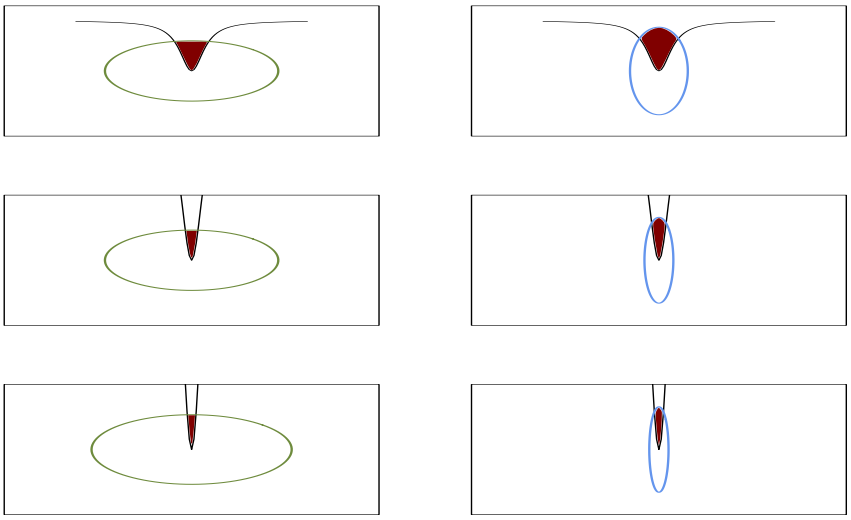

Figure 6.2: Contours of the density $\pi(x,y) \propto \exp(-x^2 - y^2 - x^2y^2)$. The left-hand plots show that a RWM with spherical covariance will find it increasingly difficult to propose values which will be accepted as the chain moves into the tails. The right-hand plots suggest that allowing the covariance to change with position might alleviate this issue.

rectangles on the upper-half plane on $\mathbb{R}^2$ (starting at $y_2 = 1$), each centred on the vertical axis, with height one and with each successive triangle a third of the width and depth of the previous. Intuitively, the density is an ever narrowing staircase, as shown in Figure 6.3.

For simplicity here we take the Random Walk Metropolis proposal as simply a uniform distribution on the disc of radius one about the current point, so $Q_R(x,A) = \mu^L(A \cap S_x)/\mu^L(S_x)$, where $S_x := \{y \in \mathbb{R}^2; |y-x| \leq 1\}$. To imitate the changing covariance in the PDRWM, we take as a proposal a uniform distribution over an ellipse for which the width is $3^{1-\lfloor x_2 \rfloor}$ if the current position is $x = (x_1, x_2) \in \mathbb{R}^2$, so $Q_P(x,A) = \mu^L(A \cap E_x)/\mu^L(E_x)$, where $E_x = \{y \in \mathbb{R}^2 : \ 3^{2(1-\lfloor x_2 \rfloor)}(y_1 - x_1)^2 + (y_2 - x_2)^2 \leq 1\}$. For these choices many of the calculations required in this section reduce to calculating areas of rectangles and ellipses.

Proposition 6.6. *The Metropolis–Hastings algorithm with proposal Q_R does not produce a geometrically ergodic Markov chain when $\pi(x) = \square(x)$.*

Proof: It is sufficient to construct a sequence of points $x_p \in \mathbb{R}^2$ such that $|x_p| \to \infty$ as $p \to \infty$, and

show that $r(x_p) \to 1$. Take $x_p = (0, p)$ for $p \in \mathbb{N}$. In this case $r(x_p)$ is bounded below by one minus the area of the rectangles that x_p is on the boundary of divided by the area of the circle $|S_x| = \pi$. So we have

$$r(x_p) \geq 1 - \left(\frac{1}{3^{p-2}\pi} + \frac{1}{3^{p-1}\pi} \right) \to 1$$

as $p \to \infty$, as required. ■

The approach makes it clear that reducing the area of an ellipse at the same rate as the area of the rectangles will remove this issue. The next result confirms this intuition.

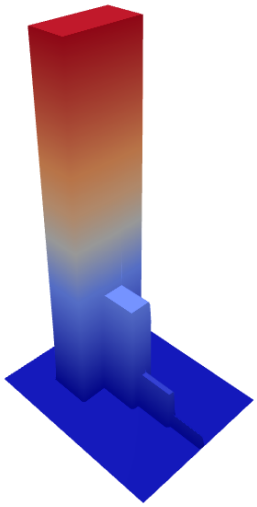

Figure 6.3: The rectangle density.

Proposition 6.7. *The Metropolis–Hastings algorithm with proposal Q_P produces a geometrically ergodic Markov chain when $\pi(x) = \square(x)$, from π-almost any starting point.*

Proof: We can take as a small set $C = \{y \in \mathbb{R}^2; 1 \leq y_i \leq 2\}$, i.e. the largest rectangle on the contour plot. Outside of this set, we show that the chain behaves in the vertical coordinate as a random walk with inwards drift, which is shown to be geometrically ergodic in Section 16.1.3 of [81]. We can therefore use the Lyapunov function $V(x) = e^{s(1 \vee |x_1| + x_2)}$, which is both coercive and only depends on the x_2 coordinate within R. Note first that $\alpha(x, y) = 1$ for any $x, y \in R \cap \{y \in \mathbf{X} : y_2 < x_2\}$. Because of this, it suffices to show that the overlap on the contour plot between the lower hemisphere of each E_x and R is larger than that between R and the upper hemisphere for any $x \in R \setminus C$, which is clearly true from inspecting Figure 6.4. This establishes that in the x_2 coordinate the chain will be of the form $y_i = y_{i-1} + \eta_i$, where η_i follows a distribution which has a negative mean, and the result follows. ■

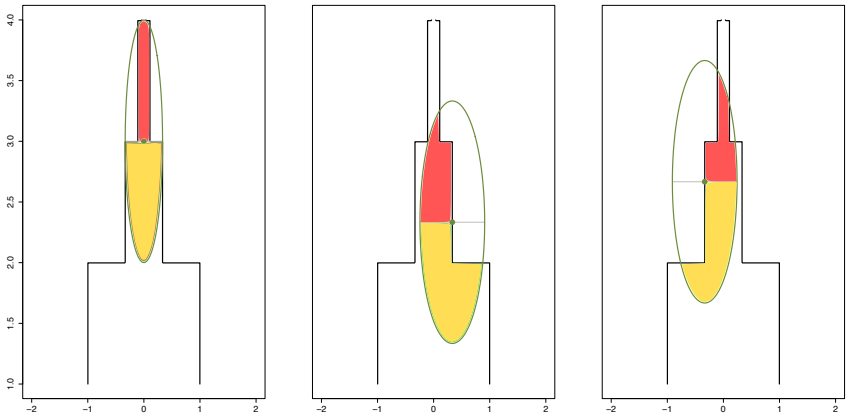

Figure 6.4: Contour plots of the rectangle density, showing the set of proposals which would be accepted if the current point is given by the green dot. The area in the lower half of the ellipse which is coloured yellow is larger than that in the upper half (shown in red), implying that on average the vertical coordinate (and hence $V(x)$) will be smaller for the next point in the chain.

6.4 Proofs

The longer proofs of results stated above are given here, so that the main idea of the paper can be grasped more easily. In each case we re-state the result and then provide a full proof.

6.4.1 Proof of Theorem 6.2

If $G^{-1}(x) = o(|x|^2)$ and $\pi(x)$ is log-concave in the tails, then the PDRWM method produces a geometrically ergodic Markov chain from π-almost any starting point. If $\pi(x)$ is subexponential for some $\beta \in (0,1)$, then choosing $G^{-1}(x) = \Theta(|x|^\gamma)$ for some $2(1-\beta) < \gamma < 2$ gives the same result.

Proof: For the log-concave case, take $V(x) = e^{s|x|}$ for some $s > 0$, and let

$$B_A := \int_A \left[\frac{V(y)}{V(x)} - 1\right] \alpha(x,y) Q(x,dy).$$

Recall from Subsection 4.2 that showing $\limsup_{|x|\to\infty} B_{(-\infty,\infty)} < 0$ is sufficient to establish the result here. We first break up $\mathbf{X}$ into $(-\infty,0] \cup (0, x - cx^{\frac{\gamma}{2}}] \cup (x - cx^{\frac{\gamma}{2}}, x + cx^{\frac{\gamma}{2}}] \cup (cx^{\frac{\gamma}{2}}, x + cx^{\gamma}] \cup (x + cx^{\gamma}, \infty)$,

and show that the integral is strictly negative on at least one of these sets, and can be made arbitrarily small as $x \to \infty$ on all others. The $-\infty$ case is analogous from the tail conditions on $\pi(x)$.

On $(\infty, 0]$, we have

$$\begin{aligned} B_{(\infty,0]} &= e^{-sx} \int_{-\infty}^{0} e^{s|y|} \alpha(x,y) Q(x,dy) - \int_{-\infty}^{0} \alpha(x,y) Q(x,dy), \\ &\leq e^{-sx} \int_{0}^{\infty} e^{sy} Q(-x,dy). \end{aligned}$$

The integral is now proportional to the moment generating function of a truncated Gaussian distribution (see Appendix D), so is given by

$$e^{-sx+x^{\gamma}hs^2/2} \left[1 - \Phi\left(x^{1-\gamma/2}/h^{1/2} - h^{1/2} s x^{\gamma/2} \right) \right].$$

A simple bound on the error function is $\sqrt{2\pi} x \Phi^c(x) < e^{-x^2/2}$ (See Appendix E), so setting $\eta = x^{1-\gamma/2}/h^{1/2} - h^{1/2} s x^{\gamma/2}$ we have

$$\begin{aligned} B_{(\infty,0]} &\leq \frac{1}{\sqrt{2\pi}} \exp\left(-2sx + \frac{hs^2}{2} x^{\gamma} - \frac{1}{2}\left(h^{-1} x^{2-\gamma} - 2sx + hs^2 x^{\gamma} \right) + \log \eta \right), \\ &= \frac{1}{\sqrt{2\pi}} \exp\left(-sx - \frac{1}{2h} x^{2-\gamma} + \log \eta \right). \end{aligned}$$

which $\to 0$ as $x \to \infty$, so we can make this arbitrarily small.

On $(0, x - cx^{\gamma/2}]$, note that $e^{s(|y|-|x|)} - 1$ is clearly negative throughout this region. So the integral is straightforwardly bounded as $B_{(0,x-cx^{\gamma/2}]} \leq 0$ for all $x \in \mathbf{X}$.

On $(x - cx^{\gamma/2}, x + cx^{\gamma/2}]$, provided $x - cx^{\gamma/2}$ is large enough that we are in the tail regime, then for any y in this region

$$\alpha(x,y) \leq \exp\left(-a(y-x) + \frac{\gamma}{2} \log\left| \frac{x}{y} \right| - \frac{1}{2h} \left[(x-y)^2 y^{-\gamma} - (x-y)^2 x^{-\gamma} \right] \right).$$

A Taylor expansion of $y^{-\gamma}$ about x gives

$$y^{-\gamma} = x^{-\gamma} - \gamma x^{-\gamma-1}(y-x) + \frac{\gamma(\gamma+1)}{2} x^{-\gamma-2}(y-x)^2 + \ldots$$

and multiplying by $(y-x)^2$ gives

$$(y-x)^2 y^{-\gamma} = \frac{(y-x)^2}{x^{\gamma}} - \gamma \frac{(y-x)^3}{x^{\gamma+1}} + \frac{\gamma(\gamma+1)}{2} \frac{(y-x)^4}{x^{\gamma+2}} + \ldots$$

If $|y-x| = cx^{\gamma/2}$ then this is:

$$\frac{c^2 x^{\gamma}}{x^{\gamma}} - \gamma \frac{c^3 x^{3\gamma/2}}{x^{\gamma+1}} + \frac{\gamma(\gamma+1)}{2} \frac{c^4 x^{2\gamma}}{x^{\gamma+2}} + \ldots$$

As $\gamma < 2$ then $3\gamma/2 < \gamma+1$, and similarly for successive terms, meaning each gets smaller as $|x| \to \infty$. So we have for large x and $y \in (x - cx^{\gamma/2}, x + cx^{\gamma/2})$

$$(y-x)^2 y^{-\gamma} \approx \frac{(y-x)^2}{x^\gamma} - \gamma \frac{(y-x)^3}{x^{\gamma+1}}. \tag{6.2}$$

Using (6.2) gives (for large enough x)

$$\alpha(x,y) \leq \exp\left(-a(y-x) + \frac{\gamma}{2}\log\left|\frac{x}{y}\right| + \frac{1}{2h}\gamma\frac{(y-x)^3}{x^{\gamma+1}}\right)$$

So we can analyse how the acceptance rate behaves. First note that for fixed $\varepsilon > 0$

$$\alpha(x, x+\varepsilon) \leq \exp\left(-a\varepsilon + \frac{\gamma}{2}\log\left|\frac{x}{x+\varepsilon}\right| + \frac{1}{2h}\gamma\frac{\varepsilon^3}{x^{\gamma+1}}\right) \to \exp(-a\varepsilon).$$

Similarly we find that the $e^{-a\varepsilon}$ term will dominate for any ε for which $\varepsilon^3/x^{\gamma+1} \to 0$, i.e. any $\varepsilon = o(x^{(\gamma+1)/3})$. If $\gamma < 2$ then $\varepsilon = cx^{\gamma/2}$ satisfies this condition. So for any $y > x$ in this region we can choose an x such that

$$\alpha(x,y) \leq \exp\left(-a(y-x) + \delta_x\right),$$

where δ_x can be made arbitrarily small in this region by choosing a large enough x. For the case $y < x$ here we have (for any fixed $\varepsilon > 0$)

$$\alpha(x, x-\varepsilon) \leq \exp\left(a\varepsilon + \frac{\gamma}{2}\log\left|\frac{x}{x-\varepsilon}\right| - \frac{1}{2h}\gamma\frac{\varepsilon^3}{x^{\gamma+1}}\right) \to \exp(a\varepsilon).$$

So by a similar argument we have $\alpha(x,y) > 1$ here for large x, as the exponential term will dominate. Combining these results we can write

$$\begin{aligned} B_{(x-cx^{\gamma/2}, x+cx^{\gamma/2}]} &= \int_0^{cx^{\gamma/2}} \left[e^{(s-a)z+\delta_z} - e^{-az+\delta_z} + e^{-sz} - 1\right] q_x(dz), \\ &= -\int_0^{cx^{\gamma/2}} (1-e^{-sz})(1-e^{(s-a)z+\delta_z}) q_x(dz), \end{aligned}$$

which will be strictly negative for large enough x provided $s < a$, where $q_x(\cdot)$ denotes a zero mean Gaussian distribution with the same variance as $Q(x,\cdot)$.

On $(x + cx^{\gamma/2}, x + cx^{\gamma}]$ we can upper bound the acceptance rate as

$$\alpha(x,y) \leq \frac{\pi(y)}{\pi(x)} \exp\left(\frac{1}{2}\log\frac{|G(y)|}{|G(x)|} + \frac{G(x)}{2h}(x-y)^2\right)$$

If $y \geq x$ and $x > x_0$ then we have

$$\alpha(x,y) \leq \exp\left(-a(|y| - |x|) + \frac{1}{2h}\frac{(x-y)^2}{x^\gamma}\right).$$

For $|y-x| = cx^{\eta}$ this becomes

$$\alpha(x,y) \leq \exp\left(-acx^{\eta} + \frac{c^2}{2h}x^{2\eta-\gamma}\right)$$

So provided $\gamma > \eta$ the e^{-a} term will dominate for large x. In the equality case we have

$$\alpha(x,y) \leq \exp\left(\left(\frac{c^2}{2h} - a\right)cx^{\gamma}\right),$$

so provided we choose c such that $a > c^2/2h$ then the acceptance rate will also decay exponentially. Because of this we have

$$\begin{aligned} B_{(x+cx^{\gamma/2},x+cx^{\gamma}]} &\leq \int_{A_4} e^{s(y-x)}\alpha(x,y)Q(x,dy), \\ &\leq e^{(c^2/2h+s-a)cx^{\gamma/2}}Q(x,(x+cx^{\gamma/2},x+cx^{\gamma}]), \end{aligned}$$

so provided $a > c^2/2h + s$ then this term can be made arbitrarily small.

On $(x+cx^{\gamma},\infty)$ using the same properties of truncated Gaussians and error function bounds we have

$$\begin{aligned} B_{(x+cx^{\gamma},\infty)} &\leq e^{-sx}\int_{x+cx^{\gamma}}^{\infty} e^{sy}Q(x,dy), \\ &= e^{s^2x^{\gamma}/2}\Phi^{c}((c-s)x^{\gamma}) \leq \exp\left(\frac{-c(c-2s)}{2}x^{\gamma}\right), \end{aligned}$$

which can be made arbitrarily small provided $c > 2s$.

For the subexponential case, the proof is similar. Take $V(x) = e^{s|x|^{\beta}}$, and divide $\mathbf{X}$ up into the same regions. Outside of $(x-x^{\gamma/2},x+x^{\gamma/2}]$ the same arguments show that the integral can be made arbitrarily small. On this set, note that in the tails.

$$(x+cx^{\frac{\eta}{2}})^{\beta} - x^{\beta} = \beta cx^{\frac{\eta}{2}+\beta-1} + \frac{\beta(\beta-1)c^2}{2}x^{\eta+\beta-2} + \ldots$$

For $y-x = cx^{\eta/2}$, then for $\eta/2 < 1-\beta$ this becomes negligible, otherwise it will grow as x does. So in this case we further divide the typical set into $(x,x+cx^{1-\beta}] \cup (x+cx^{1-\beta},x+cx^{\gamma/2})$. On $(x-cx^{1-\beta},x+cx^{1-\beta})$ the integral is bounded above by $e^{-c_1}Q(x,(x-cx^{1-\beta},x+cx^{1-\beta})) \to 0$, for some suitably chosen $c_1 > 0$. On $(x-cx^{\gamma/2},x-cx^{1-\beta}] \cup (x+cx^{1-\beta},x+cx^{\gamma/2}]$ then for $y > x$ we have $\alpha(x,y) \leq e^{-c_2(y^{\beta}-x^{\beta})}$, so we can use the same argument as in the the log-concave case to show that the integral will be strictly negative in the limit.

■

6.4.2 Proof of Theorem 6.3

If $G^{-1}(x) = \Theta(|x|^2)$, then there is a $h_0 > 0$ such that for a step-size $h \in (0, h_0)$ the PDRWM method produces a geometrically ergodic Markov chain from π-almost any starting point, provided $\pi(x) \leq |x|^{-p}$ in the tails for some $p > 1$.

Proof: Here a typical proposal will be $y = x \pm \xi\sqrt{h}x$ for x sufficiently large, meaning $|x-y| = \xi\sqrt{h}x$, with $\xi \sim N(0,1)$. For now we assume both x and y are in the tail regime, meaning $G(y) \propto y^{-2}$ and similarly for $G(x)$ (we make this concrete later). We can also take $\pi(y)/\pi(x) = x^p/y^p$ here.

For $y = (1+\xi\sqrt{h})x$ then in the tails the acceptance rate becomes

$$\alpha(x,y) = 1 \wedge \frac{1}{(1+\xi\sqrt{h})^{p+1}} \exp\left(\frac{\xi^3\sqrt{h}}{2}\left[\frac{2+\xi\sqrt{h}}{(1+\xi\sqrt{h})^2}\right]\right),$$

which is completely independent of x.

Take $V(x) = 1 \vee |x|^s$, for some $s < 1$ which is suitably small that $\int V(y)\pi(dy) < \infty$, together with an extra restriction which we specify later. Then $V(y)/V(x)$ becomes independent of x also. The integral of interest can now be re-written in terms of ξ, with $\mu^G(\cdot)$ a standard Gaussian measure, $\phi(\xi)$ its density, and $\alpha_h(\xi)$ the acceptance rate. So in most of the regions we consider we can choose x large enough that the integral in question is

$$\int \left[|1+\xi\sqrt{h}|^s - 1\right] \alpha_h(\xi)\mu^G(d\xi). \tag{6.3}$$

We therefore need to show that this integral is strictly negative for h small enough, and take care of the values of y which may not fall into this region.

Using the same shorthand B_A as in the proof of the previous Theorem, here we divide $\mathbf{X}$ into

$$\begin{aligned} B_{(\infty,\infty)} &= B_{(-\infty,-2h^{-1/2})} + B_{(-2h^{-1/2},-\delta h^{-1/4})} + B_{(-\delta h^{-1/4},\delta h^{-1/4})} + B_{(\delta h^{-1/4},\infty)}, \\ &= B_{H_1} + B_{H_2} + B_{H_3} + B_{H_4}. \end{aligned}$$

It is clear that all of these integrals can be made arbitrarily close to zero by making h small enough. The goal is to show that $B_{(\infty,\infty)} < 0$ for all $h \in (0, h_0)$. We proceed by finding the order of h of each B_{H_i}.

On $H_1 = (-\infty, -2h^{-1/2})$ we have

$$B_{H_1} \leq \frac{1}{\sqrt{2\pi}} \int_{H_1} \left[|1+\xi\sqrt{h}|^s - 1\right] \exp\left(-\frac{\xi^2}{2}\right) d\xi$$

Use the change of variables $\gamma = 1+\xi\sqrt{h}$ gives

$$B_{H_1} \leq \int_{-\infty}^{-1} [|\gamma|^s - 1]\,\mu^G(d\gamma) = \int_1^{\infty} (\eta^s - 1)\mu^G(d\eta) < \int_1^{\infty} \eta\mu^G(d\eta),$$

with $\eta \sim N(-1,h)$, as $s<1$. Using results for truncated Gaussians, we have

$$\begin{aligned}\int_1^{\infty} \eta\mu^G(d\eta) &= -\Phi\left(-\frac{2}{\sqrt{h}}\right) + \sqrt{h}\phi\left(\frac{2}{\sqrt{h}}\right)\frac{\Phi\left(-\frac{2}{\sqrt{h}}\right)}{1-\Phi\left(\frac{2}{\sqrt{h}}\right)},\\ &= -\Phi^c\left(\frac{2}{\sqrt{h}}\right) + \sqrt{h}\phi\left(\frac{2}{\sqrt{h}}\right).\end{aligned}$$

The lower bound on Φ^c from Appendix E gives

$$B_{H_1} \leq \frac{2+h}{4+h}\sqrt{\frac{h}{2\pi}}\exp\left(-\frac{2}{h}\right).$$

On $H_2 = (-2h^{-1/2}, -\delta h^{-1/4})$, the function $\left[|1+\xi\sqrt{h}|^s - 1\right]$ is negative, so this integral is trivially bounded as ≤ 0 for any h. Note that this is the entire set of y's for which (6.3) is not the correct integral.

On $H_3 = (-\delta h^{-1/4}, \delta h^{-1/4})$ recall that the acceptance probability is

$$\alpha_h(\xi) = \exp\left(-(p+1)\log(1+\xi\sqrt{h}) + \frac{\xi^3 h}{2}\left[\frac{2+\xi\sqrt{h}}{(1+\xi\sqrt{h})^2}\right]\right)$$

For any $\xi > 0$ we have

$$\frac{2+\xi\sqrt{h}}{(1+\xi\sqrt{h})^2} < \frac{2(1+\xi\sqrt{h})}{(1+\xi\sqrt{h})^2} < 2, \text{ so } \frac{\xi^3 h}{2}\left[\frac{2+\xi\sqrt{h}}{(1+\xi\sqrt{h})^2}\right] < \xi^3 h,$$

meaning

$$\alpha_h(\xi) < \exp\left(-(p+1)\log(1+\xi\sqrt{h}) + \xi^3 h\right).$$

We would like to write this as $(1+\xi\sqrt{h})^{-a}$ for some $a>0$. If $\delta h^{\frac{1}{4}} < 1$ we can use a Taylor expansion with remainder $\log(1+x) = x - x^2/2 + r^3/3$ for some $r \in (0,x)$ to get the bound $x - x^2/2 \leq \log(1+x)$ for $0 \leq x < 1$. For any $b < p+1$ then

$$b\log(1+\xi\sqrt{h}) > b\left(\xi\sqrt{h} - \frac{\xi^2 h}{2}\right) > \frac{b\xi\sqrt{h}}{2} > \xi^3 h \text{ for } \xi \in (0, \delta h^{-\frac{1}{4}}),\ \delta < \sqrt{\frac{b}{2}}.$$

So provided δ is chosen in this way then $\exists a > 0$ such that $\alpha_h(\xi) \leq (1+\xi\sqrt{h})^{-a}$ for $\xi \in (0, \delta h^{-\frac{1}{4}})$ and $\alpha = 1$ for $\xi \in (-\delta h^{-\frac{1}{4}}, 0)$ (by simply reversing the signs in the above inequalities). Now the

integral of interest can be written

$$B_{H_3} \leq \int_0^{\delta h^{-\frac{1}{4}}} \left[(1+\xi\sqrt{h})^{(s-a)} - (1+\xi\sqrt{h})^{-a} + (1-\xi\sqrt{h})^{s} - 1\right] \mu^G(d\xi).$$

So we need to bound

$$\int (1+\xi\sqrt{h})^{s-a}\mu^G(d\xi) - \int (1+\xi\sqrt{h})^{-a}\mu^G(d\xi) + \int (1-\xi\sqrt{h})^{s}\mu^G(d\xi) - \frac{1}{2}\Phi(\delta h^{-\frac{1}{4}}).$$

Upper and lower bounds for $g(\xi) = (1+\xi\sqrt{h})^{-a}$ on $(0, \delta h^{-\frac{1}{4}})$ are

$$\begin{aligned} g_u(\xi) &= m_u(a)\xi + 1, \;\; m_u(a) = \frac{h^{\frac{1}{4}}}{\delta}\left[(1+\delta h^{\frac{1}{4}})^{-a} - 1\right], \\ g_l(\xi) &= m_l(a)\xi + 1, \;\; m_l(a) = -a\sqrt{h}. \end{aligned}$$

The first is a straight line through $g(\delta h^{-\frac{1}{4}})$ and $g(0) = 1$, the second is the straight line through $g(0) = 1$ with gradient $g'(0)$ (as the function is convex). This gives upper and lower bounds for the first two integrals as

$$m_u(a-s)\Psi_h + \Phi(\delta h^{-\frac{1}{4}}) - \frac{1}{2}, \;\text{ and }\; m_l(a)\Psi_h + \Phi(-\delta h^{\frac{1}{4}}) - \frac{1}{2}.$$

where $\Psi_h = \phi(\delta h^{-\frac{1}{4}}) - 1/\sqrt{2\pi} < 0$. We can construct a similar Taylor Series upper bound for $(1-\xi\sqrt{h})^s$ as a straight line with gradient $m_u^* = -s\sqrt{h}$ (as this function is concave), meaning the total bound of interest is

$$\begin{aligned} B_{H_3} &\leq (m_u(a-s) - m_l(a) + m_u^*)\Psi_h, \\ &= \left((a-s)\sqrt{h} + \frac{h^{\frac{1}{4}}}{\delta}\left((1+\delta h^{\frac{1}{4}})^{s-a} - 1\right)\right)\Psi_h, \\ &= C_{H_3}\exp\left(-\frac{\delta^2}{2\sqrt{h}}\right) - C_{H_3}, \end{aligned}$$

where $C_{H_3} = (a-s)\sqrt{h} + \frac{h^{\frac{1}{4}}}{\delta}\left((1+\delta h^{\frac{1}{4}})^{s-a} - 1\right)$. To see that C_{H_3} is positive, we can Taylor expand $(1+\delta h^{1/4})^{s-a}$, so that

$$\begin{aligned} C_{H_3} &= (a-s)\sqrt{h} + \frac{h^{\frac{1}{4}}}{\delta}\left((1+\delta h^{\frac{1}{4}})^{s-a} - 1\right), \\ &= (a-s)\sqrt{h} + \frac{h^{\frac{1}{4}}}{\delta}\left(-(a-s)\delta h^{\frac{1}{4}} + \frac{(s-a)(s-a-1)}{2}\delta^2 h^{1/2} + O(h^{\frac{3}{4}})\right), \\ &= \frac{(s-a)(s-a-1)}{2}\delta h^{3/4} + O(h) > 0. \end{aligned}$$

On $H_4 = (\delta h^{-1/4}, \infty)$, bounding in the same way as for H_1, we set $\gamma = 1 + \xi\sqrt{h}$, meaning $\gamma \sim N(1,h)$. Then

$$B_{H_4} \leq \int_{\delta h^{-\frac{1}{4}}}^{\infty} [|\gamma|^s - 1]\, \mu^G(d\gamma),$$

which can be re-written

$$\begin{aligned}\mathbb{E}_{\varpi}[|\gamma|^s - 1]\,\Phi^c(\delta h^{-\frac{1}{4}}) &\leq \mathbb{E}_{\varpi}[\gamma]\,\Phi^c(\delta h^{-\frac{1}{4}}), \\ &= (1 + \delta h^{\frac{1}{4}})\Phi^c(\delta h^{-\frac{1}{4}}) + \sqrt{h}\phi(\delta h^{-\frac{1}{4}}),\end{aligned}$$

where ϖ is now a truncated Gaussian distribution on $(1 + \delta h^{\frac{1}{4}}, \infty)$ with mean 1 and variance h. Using the upper bound on Φ^c gives

$$\begin{aligned}B_{H_4} &\leq (1 + \delta h^{\frac{1}{4}})\frac{1}{\sqrt{2\pi}}\frac{h^{\frac{1}{4}}}{\delta}\exp\left(-\frac{\delta^2}{2\sqrt{h}}\right) + \sqrt{\frac{h}{2\pi}}\exp\left(-\frac{\delta^2}{2\sqrt{h}}\right), \\ &= \sqrt{\frac{h^{\frac{1}{4}}}{2\pi}}\left(2h^{\frac{1}{4}} + \frac{1}{\delta}\right)\exp\left(-\frac{\delta^2}{2\sqrt{h}}\right), \\ &= C_{H_4}\exp\left(-\frac{\delta^2}{2\sqrt{h}}\right)\end{aligned}$$

Combining inequalities, we can get a very loose upper bound on the integral as

$$B_{(-\infty,\infty)} \leq (C_{H_4} + C_{H_3})\exp\left(-\frac{\delta^2}{2\sqrt{h}}\right) + C_{H_1}\exp\left(-\frac{2}{h}\right) - C_{H_3}.$$

The exponentials are the dominant terms in the first two expressions, as they shrink to zero much faster than any of the C_{H_i} terms (which still depend on h). We have already shown that C_{H_3} is $O(h)$, and in fact it is more straightforward to see that C_{H_1} and C_{H_4} are both $O(h^{\frac{1}{2}})$. Because of this, we can always choose a h small enough that the last term is arbitrarily larger than all others in the expression, meaning that the integral is strictly negative, as required.

■

6.4.3 Proof of Theorem 6.4

If $G^{-1}(x) = \omega(|x|^2)$, then the PDRWM method can never produce a geometrically ergodic Markov chain provided $\pi(y) \leq \pi(x)$ for all $|Y| \geq |x| \geq L$, for some $L < \infty$.

Proof: The goal is to show

$$\limsup \int \alpha(x,y)Q(x,dy) = 0.$$

The general strategy will be to find some set

$$A_{x,\varepsilon} := \{y \in \mathbf{X} : \alpha(x,y) \geq \varepsilon\}.$$

In words, a set which shows the potential candidate moves which have a non-negligible probability of acceptance. We will then establish that $Q(x,A_{x,\varepsilon}) \to 0$ as $x \to \infty$, for any $\varepsilon > 0$.

First recall that for the algorithm in general the acceptance probability for a proposal y is

$$\alpha(x,y) = \frac{\pi(y)|G(y)|^{\frac{1}{2}}}{\pi(x)|G(x)|^{\frac{1}{2}}} \exp\left(-\frac{1}{2h}(y-x)^2[G(y)-G(x)]\right).$$

If $G(x) = \Theta(|x|^{-\gamma})$, then for large enough x and y the acceptance probability is

$$\alpha(x,y) = 1 \wedge \frac{\pi(y)}{\pi(x)} \left(\frac{|x|}{|y|}\right)^{\frac{\gamma}{2}} \exp\left(-\frac{c}{2h}(x-y)^2\left[\frac{1}{|y|^\gamma} - \frac{1}{|x|^\gamma}\right]\right).$$

As each $Q(x,\cdot)$ is a Gaussian distribution, we consider a 'typical set' to be

$$T_x = \left(x - 2\sqrt{h}x^{\gamma/2}, x + 2\sqrt{h}x^{\gamma/2}\right).$$

For any x, $Q(x,T_x) \approx 0.96$. If we can show that i) for large enough x, $A_{x,\varepsilon} \subset T_x$, and ii) the ratio $Q(x,A_{x,\varepsilon})/Q(x,T_x) \to 0$ then we will have established the result.

First we note that for $|y|$ larger than $x > L$ then the assumptions directly imply that $\pi(y)/\pi(x) \leq 1$, so we can say

$$\alpha(x,y) \leq \left(\frac{x}{|y|}\right)^{\frac{\gamma}{2}} \exp\left(-\frac{c}{2h}\left[\frac{(x-y)^2}{|y|^\gamma} - \frac{(x-y)^2}{x^\gamma}\right]\right).$$

Since if $y = x$ then $\alpha(x,y) = 1$, we will only concern ourselves with $|y| > |x|$. In effect we are now considering the set $A_{x,\varepsilon} \cup (-x,x)$, but since this is strictly larger than $A_{x,\varepsilon}$ it will give us the result. For $y > x$, if we write $y = x+z$ for some $z > 0$ (and do similar in the other tail), we can see that

$$\alpha(x,x+z) \leq \left(\frac{x}{x+z}\right)^{\frac{\gamma}{2}} \exp\left(-\frac{cz^2}{2h(x+z)^\gamma} + \frac{cz^2}{2hx^\gamma}\right).$$

As $x \to \infty$, the first term on the right-hand side will tend to something greater than zero for $z = O(x)$ and decay to zero for the set of z's that grow at a larger rate than x . Inside the exponential, the term $cz^2/2h(x+z)^\gamma \to 0$ for any z as x grows. The last term $cz^2/2hx^\gamma$ will only increase with x for the set of z's that grow at a faster rate than $x^{\gamma/2}$. If we denote this set of 'extreme' values for y which would be accepted as $E_{x,\varepsilon} = A_{x,\varepsilon} \cap T_x^c$, then it is clear that $Q(x,E_{x,\varepsilon}) \to 0$ for any $\varepsilon > 0$, as $E_{x,\varepsilon} \sim (-\infty, -x^{\gamma/2+\delta}) \cup (x^{\gamma/2+\delta}, \infty)$ for some $\delta > 0$, and this set will be sent deeper and deeper into the tails of $Q(x,\cdot)$ as $|x|$ grows.

So now we can focus on $A_{x,\varepsilon} \cap T_x$, or equivalently consider the set of possible z values in $(-2x^{\gamma/2}, 0) \cup (0, 2x^{\gamma/2})$. For any of these the dominant term in $\alpha(x, x+z)$ will be $(x/(x+z))^{\gamma/2}$, so the acceptance rate will be strictly decreasing in z on this set. Hence we need only examine the boundary points, $y = x + 2\sqrt{h}x^{\gamma/2}$ and $y = x - 2\sqrt{h}x^{\gamma/2}$, and show that these both decay to zero as $x \to \infty$.

For $y = x + 2\sqrt{h}x^{\gamma/2}$ the acceptance rate becomes

$$\begin{aligned}
\alpha(x,y) &\leq \left(\frac{x}{x+2\sqrt{h}x^{\gamma/2}}\right)^{\gamma/2} \exp\left(-\frac{c}{2h}\left[\frac{4\sqrt{h}x^{\gamma}}{|x+2\sqrt{h}x^{\gamma/2}|^{\gamma}} - 4\sqrt{h}\right]\right), \\
&\leq \left(\frac{x}{x+2\sqrt{h}x^{\gamma/2}}\right)^{\gamma/2} \exp\left(\frac{2c}{\sqrt{h}}\right), \\
&\to 0.
\end{aligned}$$

And for $y = x - 2\sqrt{h}x^{\gamma/2}$, noting that for large x $|x - 2\sqrt{h}x^{\gamma/2}| > \sqrt{h}x^{\gamma/2}$, we have

$$\begin{aligned}
\alpha(x,y) &\leq \left(\frac{x}{\sqrt{h}x^{\gamma/2}}\right)^{\gamma/2} \exp\left(\frac{2c}{\sqrt{h}}\right) \exp\left(-\frac{c}{2h}\left[\frac{4\sqrt{h}x^{\gamma}}{x^{\gamma^2/2}}\right]\right), \\
&\leq \left(\frac{x}{\sqrt{h}x^{\gamma/2}}\right)^{\gamma/2} \exp\left(\frac{2c}{\sqrt{h}}\right), \\
&\to 0.
\end{aligned}$$

■

6.5 Discussion

In this chapter we have analysed the ergodic behaviour of a Metropolis-Hastings method with proposal kernel $Q(x,\cdot) = N(x, hG^{-1}(x))$. In one dimension we have characterised the behaviour in terms of growth conditions on $G^{-1}(x)$ and tail conditions on the target distribution, and some cases in higher dimensions have also been discussed. The goal was to understand whether generalising an existing Metropolis–Hastings method by allowing the proposal covariance to change with position can alter the ergodic properties of the sampler. We can confirm that this is indeed possible, either for better or worse, depending on the choice of covariance. The key points for practitioners are i) lack of sufficient care in the design of $G^{-1}(x)$ can have severe consequences (as in Theorem 6.4), and ii) careful choice of $G^{-1}(x)$ can have much more beneficial ones, particularly in higher dimensions, as evidenced by the 'rectangle' density example.

We feel that such results can also offer insight into similar generalisations of different Metropolis–

Hastings algorithms (e.g. [43, 130]). For example, it seems intuitive that any method in which the variance grows at a faster than quadratic rate in the tails is unlikely to produce a geometrically ergodic chain. There are connections between the PDRWM and some extensions of the Metropolis-adjusted Langevin algorithm [130], the ergodicity properties of which are discussed in [62]. The key difference between the schemes is the inclusion of the drift term $G^{-1}(x)\nabla \log \pi(x)/2$ in the latter. It is this term which in the main governs the behaviour of the sampler, which is why the behaviour of the PDRWM is different to this scheme (note that gradients are required for all variants, unlike in the PDRWM).

We can apply the general results to the specific variants discussed in Section 6.1. Provided sensible choices of regions/weights, and diminishing adaptation schemes are chosen, the Regional adaptive Metropolis–Hastings, Locally weighted Metropolis and Kernel-adaptive Metropolis–Hastings samplers should all satisfy $G^{-1}(x) \to \Sigma$ as $|x| \to \infty$, meaning they will inherit the ergodicity properties of the standard RWM (the behaviour in the centre of the space, however, will likely be different). In the State-dependent Metropolis method provided $b \leq 2$ (with suitable tuning in the equality case) then the sampler should also behave reasonably. Whether or not a large enough value of b would be found by a particular adaptation rule in the subexponential case is not entirely clear, and this could be an interesting direction of further study. The Tempered Langevin diffusion scheme, however, will fail to produce a geometrically ergodic Markov chain whenever the tails of $\pi(x)$ are lighter than that of a Cauchy distribution. In the case of Gaussian tails, for example, $G^{-1}(x) = e^{x^2/2}I$. To allow reasonable tail exploration, two pragmatic options would be to upper bound $G^{-1}(x)$ manually or use this scheme in conjunction with another, as there is evidence that the sampler can perform favourably when exploring the centre of a distribution [108]. None of the specific variants discussed here are able to mimic the local curvature of $\pi(x)$ in the tails, so as to enjoy the favourable behaviour exemplified in Proposition 6.7. This is possible using Hessian information as in [43], though should also be possible in cases where this isn't available using appropriate surrogates, at least in some cases.

It is reasonable to ask whether exploring the tails of a distribution adequately is always necessary. If the functions a practitioner is interested in estimating are such that $\int_C f(x)\tilde{\pi}(dx) \approx \int f(x)\pi(dx)$, where $\tilde{\pi}(\cdot)$ is the target restricted to the centre of the space C, then perhaps this is not so important. Some results in this direction are given in [15]. If this approach is taken, however, whether or not a sampler will perform appropriately becomes a considerably more problem-dependent question.

Geometric ergodicity, whilst by no means guaranteeing sensible estimators in the non-asymptotic context, does give steps towards this in some *generality*, through (3.15). As mentioned earlier, it also appears to have other favourable consequences [63, 77]. As such, we feel it is a property worth establishing.

Chapter 7

Stability of Hamiltonian Monte Carlo

This chapter is based on joint work with Michael Betancourt, Simon Byrne and Mark Girolami. It was also aided by useful discussions with Alexandros Beskos and Gareth Roberts.

The Hamiltonian Monte Carlo algorithm was introduced in Section 4.1.4. There we mentioned that comparatively little is understood rigorously about the method. In this chapter we deconstruct the algorithm, and begin to analyse its mixing properties.

We first discuss how the method can be viewed *marginally* on position space $\mathbf{X}$ in Section 7.1. We then use a simple argument to show φ-irreducibility, before giving some conditions under which the algorithm will and will not be geometrically ergodic. Some of the results presented here are confined to one dimension, and the positive geometric ergodicity results are specifically for the one-dimensional class of targets with densities of the form

$$\pi(x) \propto \exp\left(-\beta^{-1}|x|^{\beta}\right),$$

for some $\beta > 0$. By varying the choice of β this class encompasses a wide variety of tail behaviours. The special cases $\beta = 1$ and $\beta = 2$ correspond to the Laplace and Gaussian densities respectively, while $\beta \geq 1$ is needed for log-concavity. We refer to this class as the *one-dimensional exponential family*. Figure 7.1 shows contour plots of the resulting joint densities of (x, p) for different choices of β, with $p \sim N(0,1)$. We discuss how to generalise these results in Section 7.5. Analysis is restricted here to the case where the Hamiltonian is separable, meaning the momentum variance $G(x) = M$ is

independent of the current position x. Throughout we set $M = I$, for ease of exposition but without loss of generality.

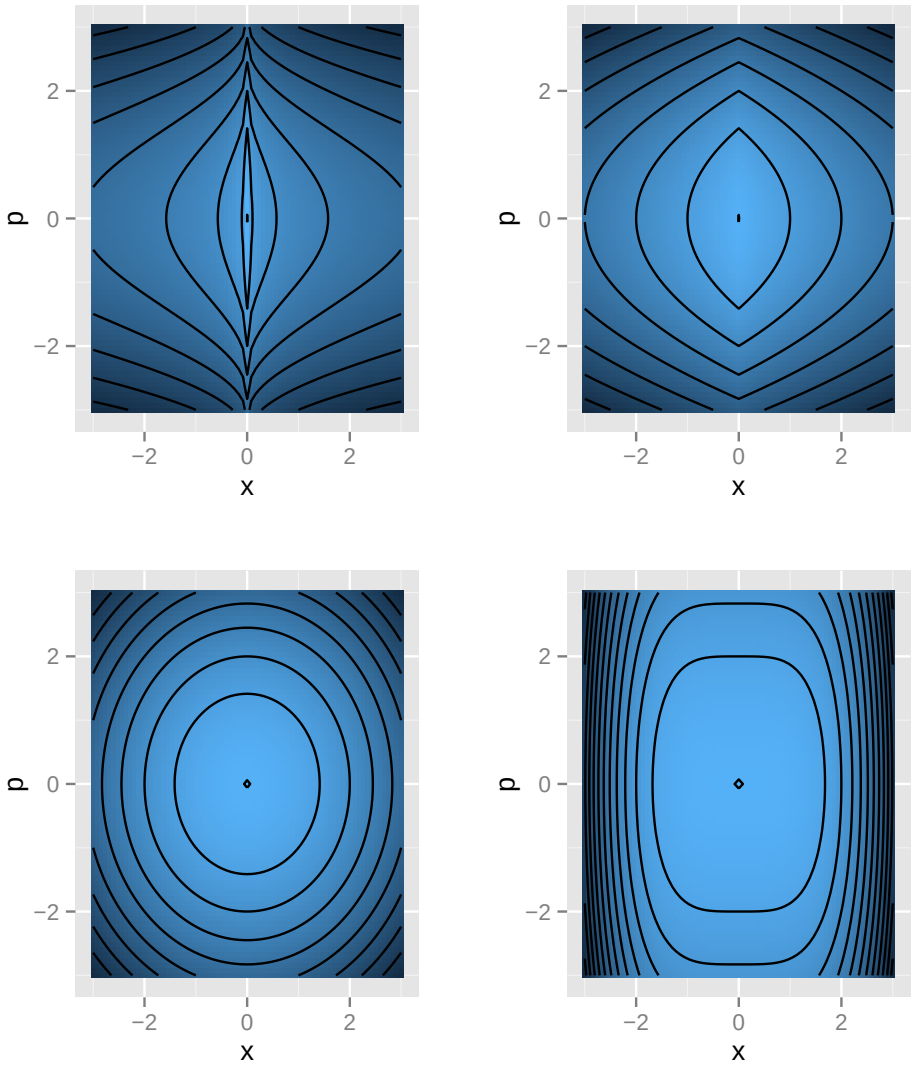

Figure 7.1: Contour plots of the joint densities $e^{-H(x,p)}$ for Hamiltonians of the form $H(x,p) = \beta^{-1}|x|^{\beta} + p^2/2$. Clockwise from the top left the parameter values are $\beta = 0.4, 1, 4$ and 2 respectively.

Some additional notation is used in this chapter. Let $\nu_{x_t,p_t}(ds) = \zeta_{x_t,p_t}^{-1} \mathbb{1}_{[0,\zeta_{x_t,p_t}]} ds$ be the Uniform distribution between 0 and ζ_{x_t,p_t}. We write $U(x_t) = -\log \pi(x_t)$ as the *potential energy*, and $K(p_t) =$

$p_t^T M^{-1} p_t/2$ as the *kinetic energy*, meaning the Hamiltonian takes the form.

$$H(x_t, p_t) = U(x_t) + K(p_t).$$

We write $\text{sgn}(x) := x/|x|$ (for $x \in \mathbb{R}$), and occasionally use the Newtonian notation $\dot{x}_t := dx_t/dt$ for time derivatives.

7.1 Constructing the marginal chain

Recall that in Hamiltonian Monte Carlo an approximation to the measure-preserving *Hamiltonian flow* is constructed using the *leapfrog* integrator. This approximate flow for some number of leapfrog steps L and integration step-size ε is used to generate a Metropolis–Hastings proposal. If the current point is $x = x_t$ then the proposal is denoted $x_{t+L\varepsilon} = \eta^x_{L\varepsilon}(x_t, p_t)$, where $p_t \sim N(0, M)$ is an auxiliary *momentum* variable. This proposal is then accepted with probability $\alpha = 1 \wedge e^{H(x_t,p_t) - H(\eta_{L\varepsilon}(x_t,p_t))}$, where $H : \mathbf{X} \times \mathbf{X} \to [0,\infty)$ is the *Hamitonian function*. If we take the current point as $x = x_t$ and set $p_t \sim N(0, I)$, then a single leapfrog iteration is given by

$$\begin{aligned} p_{t+\varepsilon/2} &= p_t + \varepsilon \nabla \log \pi(x_t)/2, \\ x_{t+\varepsilon} &= x_t + \varepsilon p_{t+\varepsilon/2}, \\ p_{t+\varepsilon} &= p_{t+\varepsilon/2} + \varepsilon \nabla \log \pi(x_{t+\varepsilon})/2. \end{aligned}$$

This transition can be *marginalised*, and instead written as

$$x_{t+\varepsilon} = x_t + \varepsilon^2 \nabla \log \pi(x_t)/2 + \varepsilon p_t \tag{7.1}$$

$$p_{t+\varepsilon} = p_t + \varepsilon \nabla \log \pi(x_t)/2 + \varepsilon \nabla \log \pi(x_{t+\varepsilon})/2. \tag{7.2}$$

From (7.1), it is clear that the proposal kernel for HMC using a single leap-frog step is in fact equivalent to that used in MALA (as has previously been noted, e.g. [43]). To see that the acceptance rates are also equal, denote $c(x) := x + \varepsilon^2 \nabla \log \pi(x)/2$ and $y = c(x) + \varepsilon p_t$, so that in the MALA case we have

$$\begin{aligned} \log \frac{q(x|y)}{q(y|x)} &= \frac{1}{2\varepsilon^2} \left(|y - c(x)|^2 - |x - c(y)|^2 \right), \\ &= \frac{1}{2\varepsilon^2} \left(|\varepsilon p_t|^2 - |\varepsilon \left(\varepsilon \nabla \log \pi(x)/2 + \varepsilon \nabla \log \pi(c(x) + \varepsilon p_t)/2 + p_t \right)|^2 \right) \\ &= \frac{1}{2} \left(|p_t|^2 - |p_{t+\varepsilon}|^2 \right), \end{aligned}$$

meaning that

$$\frac{\pi(y)q(x|y)}{\pi(x)q(x|y)} = \exp\left(H(x,p) - H(\eta_\varepsilon(x,p))\right),$$

where the left-hand side denotes the MALA acceptance rate and the right-hand side that used in HMC.

We can employ the same marginalisation after more than one leapfrog step. After two steps the marginal transition is

$$\begin{aligned} x_{t+2\varepsilon} &= x_t + \varepsilon^2 \nabla \log \pi(x_t) + \varepsilon^2 \nabla \log \pi(x_{t+\varepsilon}) + 2\varepsilon p_t, \\ p_{t+2\varepsilon} &= p_t + \varepsilon \nabla \log \pi(x_t)/2 + \varepsilon \nabla \log \pi(x_{t+\varepsilon}) + \varepsilon \nabla \log \pi(x_{t+2\varepsilon})/2. \end{aligned}$$

From this we can see one reason why HMC is challenging to analyse. After a single leapfrog step the HMC proposal reduces to the current point x_t plus a deterministic step $\varepsilon^2 \nabla \log \pi(x_t)/2$, combined with some additive Gaussian noise εp_t. Hence the proposal $Q(x,\cdot) = N(x + \varepsilon^2 \nabla \log \pi(x)/2, \varepsilon^2 I)$. After another leapfrog step, however, the proposal now involves the term $\nabla \log \pi(x_{t+\varepsilon}) = \nabla \log \pi(x_t + \varepsilon^2 \nabla \log \pi(x_t)/2 + \varepsilon p_t)$. If the map $\nabla \log \pi : \mathbf{X} \to \mathbf{X}$ is nonlinear, then this term will be a nonlinear transformation of the Gaussian p_t, so will no longer itself be Gaussian. So whenever more than one leapfrog step is taken, the HMC transition kernel often becomes intractable.

After L leapfrog steps, the marginal transitions are

$$x_{t+L\varepsilon} = x_t + L\varepsilon^2 \nabla \log \pi(x_t)/2 + \varepsilon^2 \sum_{i=1}^{L-1} (L-i) \nabla \log \pi(x_{t+i\varepsilon}) + L\varepsilon p_t, \tag{7.3}$$

$$p_{t+L\varepsilon} = p_t + \varepsilon \nabla \log \pi(x_t)/2 + \varepsilon \sum_{i=1}^{L-1} \nabla \log \pi(x_{t+i\varepsilon}) + \varepsilon \nabla \log \pi(x_{t+L\varepsilon})/2. \tag{7.4}$$

This sheds some light on the behaviour of the method, as the marginal transition (7.3) is essentially the current point combined with a sequence of gradient steps. However, the non-Gaussianity of the proposal noise still persists whenever the gradient map is nonlinear.

The acceptance rate can also be thought of marginally. Because the leapfrog method is a *symplectic* integrator, it is volume-preserving (as shown in Section 4.1.4). So the density $q(x_{t+L\varepsilon}|x_t)$ is the same as that of the momentum p_t responsible for generating $x_{t+L\varepsilon}$, which is $\propto e^{-p_t^2/2}$. Owing to the symmetry of the Gaussian distribution about zero and the reversibility of the flow, it is also true that $q(x_t|x_{t+L\varepsilon}) \propto e^{-p_{t+L\varepsilon}^2}$. So the complete method can simply be thought of as a Metropolis–Hastings method on the space $\mathbf{X}$ with proposal (7.3).

Despite the non-Gaussianity, from (7.3) and (7.4) we can immediately guess the ergodic properties of the method, following the behaviour of MALA. If $|\nabla \log \pi(x)| \to 0$ as $|x| \to \infty$, then (7.3) will reduce to $x_t + L\varepsilon p_t$ in the tails, i.e. a Random Walk Metropolis proposal. Since this condition on the gradient implies $\pi(x)$ will not be log-concave in the tails, then we can guess that the method won't produce a geometrically ergodic chain here. Similarly in the case $|\nabla \log \pi(x)|/|x| \to \infty$ as $|x| \to \infty$, it is likely that proposals will 'explode' in the tails, and almost all will be rejected, again leading to a chain which will not be geometrically ergodic. In between these two cases (when the tails of $\pi(x)$ are in between Exponential and Gaussian) it seems reasonable to assume that the sampler will behave sensibly.

Note, however, that more can be said from (7.3). The discussion so far has assumed that the number of leapfrog steps L does not depend on the current position x_t. In the heavy-tailed case, however, where the gradient becomes arbitrarily small as $|x_t|$ grows, then *increasing* the number of leapfrog steps in the proposal could result in a sampler that retains a strong drift towards the centre of the space, and is therefore much more likely to produce a geometrically ergodic Markov chain. In practice naively setting $L = L(x_t, p_t)$ may mean that the map $\eta_{L\varepsilon}$ is no longer reversible, so care would need to be taken in any such implementation to ensure that the resulting Markov chain targets the correct distribution.

These issues are explored in more detail in the next sections, where we discuss two different implementations of Hamiltonian Monte Carlo:

1. *Static* HMC, in which the number of leapfrog steps L (and hence the integration time $L\varepsilon$) is fixed
2. *Dynamic* HMC, in which $L = L(x_t, p_t)$, so that the integration time changes with position.

In the dynamic case we confine our analysis to an idealised version of the method, but also discuss practical implementations which are related to this.

7.2 Stability with fixed integration times

In the next two subsections we discuss irreducibility and ergodicity for the static version of the method.

7.2.1 φ-irreducibility

We first present a simple example that shows how the proposal transition given by (7.3) can produce a method which is not π-irreducible, and hence will not be ergodic.

Example 7.1. *Take $\pi(x) \propto e^{-x^2/2}$, meaning $\nabla \log \pi(x) = -x$, and set $L = 2$. Then the HMC proposal becomes*

$$\begin{aligned} x_{t+2\varepsilon} &= x_t - \varepsilon^2 x_t - \varepsilon^2(x_t - \varepsilon^2 x_t + \varepsilon p_t) + 2\varepsilon p_t, \\ &= x_t - \varepsilon^2 x_t - \varepsilon^2 x_t + \varepsilon^4 x_t - \varepsilon^3 p_t + 2\varepsilon p_t, \\ &= (1 - 2\varepsilon^2 + \varepsilon^4) x_t + (2\varepsilon - \varepsilon^3) p_t. \end{aligned}$$

Setting $\varepsilon = \sqrt{2}$ means $2\varepsilon - \varepsilon^3 = 0$, so that

$$x_{t+2\varepsilon} = (1 - 4 + 4) x_t = x_t.$$

With this transition, the chain does not move, the proposal kernel is simply $Q(x, \cdot) = \delta_x(\cdot)$, and hence $\{X_t\}_{t \geq 0}$ will not be π-irreducible.

Although it is in some sense trivial, the above example highlights that establishing π-irreducibility is not so straightforward here.

The example occurs in part because Hamilton flow here is *periodic*. The flow travels along the contours of equal density, so provided these contours are disjoint unions of closed curves, then the flow will travel along one such curve and eventually come back on itself. In the simple example where $\pi(x)$ is a Gaussian, meaning $H(x,p) = x^2/2 + p^2/2$, then the contours will be circles. Since the flow induced by this Hamiltonian is periodic, we can compute the *period length* ζ_{x_t, p_t} as the minimum $\zeta_{x_t, p_t} \in \mathbb{R}$ such that $\varphi_{\zeta_{x_t, p_t}}(x_t, p_t) = (x_t, p_t)$. This can be found explicitly here by calculating the length of the contour $C_{x_t, p_t} = \{(x', p') \in \mathbf{X} \times \mathbf{X} : H(x', p') = H(x_t, p_t)\}$ and the speed of the flow φ_t. The former is simply the circumference of a circle of radius $\sqrt{x^2 + p^2} = \sqrt{2H(x,p)}$. The latter is simply the Euclidean norm of Hamilton's equations, in this case

$$|J \nabla H(x,p)| = \sqrt{x^2 + p^2} = \sqrt{2H(x,p)}, \quad \text{where } J = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

Using the relation time = distance/speed, we can see that the period length in this case is

$$\zeta_{x_t,p_t} = \frac{2\pi\sqrt{2H(x,p)}}{\sqrt{2H(x,p)}} = 2\pi,$$

which is independent of the starting position (x_t, p_t). The leapfrog scheme shares this periodic behaviour [65], which is why in the above example we are able to construct a chain which simply remains at the current point. More generally, any integration time $T = L\varepsilon$ which is a factor of the period length will result in a scheme which is not π-irreducible here. Outside of the Gaussian case, the period length will depend on the current position (x_t, p_t), and hence will either slow down or speed up as x_t grows. In the next section we show that in the latter case numerical schemes will typically become unstable.

Returning to the general problem, in [19] the authors prove π-irreducibility of the HMC transition under the assumption that $\pi(x) \geq c > 0$ for any $x \in \mathbf{X}$, or equivalently that the potential energy $U(x) = -\log \pi(x)$ is bounded above, so $U(x) \leq M < \infty$. Although the proof is impressive, and holds for much more general schemes than the simple leapfrog integrator discussed here, this condition is unfortunately too restrictive for our needs. Indeed, any form of $U(x)$ which is the negative logarithm of a probability density will necessarily grow indefinitely as $|x| \to \infty$, so the condition will not hold here unless the state space $\mathbf{X}$ is compact.

Fortunately, using equation (7.3), we can actually construct a simple proof of μ^L-irreducibility under certain assumptions on $\pi(x)$, which is sufficient for our needs.

Theorem 7.2. *In the case $\mathbf{X} = \mathbb{R}^n$, if $\nabla \log \pi(x) \in C(\mathbb{R}^n)$, the set of continuous functions on $\mathbb{R}^n$, $\pi(x)$ is bounded away from 0 and ∞ on compact sets, and for every $1 \leq i \leq n$*

$$\limsup_{|x|\to\infty} \left| |x|^{-d} \frac{\partial}{\partial x_i} \log \pi(x) \right| = C \geq 0$$

as $|x| \to \infty$, for some $d \in (0,1)$, then the Hamiltonian Monte Carlo method produces a μ^L-irreducible Markov chain, and all compact sets are small.

Proof. We give the proof in the case $\mathbf{X} = \mathbb{R}$. The extension to higher dimensions is simply applying the same argument to each coordinate separately. The proof is in three stages: i) we establish that any open set $O \subset \mathbb{R}$ satisfies $P(x,O) > 0$ from any $x \in \mathbf{X}$ (note that the assumptions on $\pi(\cdot)$ imply its equivalence to Lebesgue measure), ii) we extend this to any set A for which $\mu^L(A) > 0$, showing μ^L-irreducibility, and iii) we show that this implies all compact sets are small.

For i), note that from (7.3) after L leapfrog steps the HMC proposal will be

$$x_{t+L\varepsilon} = x_t + L\varepsilon^2 \nabla \log \pi(x_t)/2 + \varepsilon^2 \sum_{i=1}^{L-1} (L-i) \nabla \log \pi(x_{t+i\varepsilon}) + L\varepsilon p_t.$$

Fix x_t and consider $x_{t+L\varepsilon} = x_{L\varepsilon}(p_t)$ as a function of p_t. The growth assumptions made for $\nabla \log \pi$ imply that $L\varepsilon p_t$ is the leading order term in $x_{t+L\varepsilon}(p_t)$, meaning $x_{t+L\varepsilon}(p_t) \to \infty$ as $p_t \to \infty$ and $x_{t+L\varepsilon}(p_t) \to -\infty$ as $p_t \to -\infty$. Since $L\varepsilon p_0$ and each $\nabla \log \pi(x_{t+i\varepsilon})$ are continuous functions then so is their sum, so by the intermediate value theorem $x_{t+L\varepsilon} : \mathbb{R} \to \mathbb{R}$, i.e. the range is the entirety of $\mathbb{R}$. Continuity for a function f implies that for any open $O \subset \mathbb{R}$, the preimage

$$f^{-1}(O) = \{y \in \mathbb{R} : f(y) \in O\}$$

is also open. Using this fact, and given that $\mathbb{P}[p_t \in f^{-1}(O)] > 0$ here, it is straightforward to see that $Q(x, O) > 0$ for any open $O \in \mathbb{R}$. The conditions on $\pi(x)$ ensure that there is a positive probability of accepting any proposed move, as in Theorem 2.2 in [110], meaning $P(x, O) > 0$ as required.

Lemma 2 in [19] shows that i) $\Rightarrow$ ii) here. Part iii) follows from Theorem 4.6. ■

For the one-dimensional exponential family of distributions, the conditions of Theorem 7.2 hold for $\beta < 2$. The result can be generalised without too much work, but as we shall see below, this is sufficient for all of the geometric ergodicity results of the next section to be valid. Note that in the cases $\beta = 1$ and $\beta = 2$ the proposal kernel in fact reduces to a Gaussian, as the function $\nabla \log \pi(x)$ is either linear or constant, meaning μ^L-irreducibility is a trivial consequence here.

7.2.2 Geometric ergodicity

As HMC is both π-invariant and aperiodic, then under the conditions of Theorem 7.2 the limiting distribution of the resulting Markov chain will be $\pi(\cdot)$. We now discuss when convergence to this limit will occur at a geometric rate in m, the number of iterations of the chain. We begin with a negative result in the case where the density $\pi(x)$ has heavy tails.

Proposition 7.3. *For a fixed number of steps L, step-size ε and mass matrix M (we take $M = I$ here for brevity), if $|\nabla \log \pi(x)| < C$ for every $x \in \mathbf{X}$, then the Hamiltonian Monte Carlo method can* only *produce a geometrically ergodic Markov chain if $\mathbb{E}_\pi[e^{s|x|}] < \infty$ for some $s > 0$.*

Proof: Recall from Section 4.2 that if for any $\xi > 0$ then we can choose a $\delta > 0$ (independent of

x) such that $Q(x, B_\delta(x)) > 1 - \xi$, then a Metropolis–Hastings algorithm with proposal Q can only produce a geometrically ergodic chain if $\mathbb{E}_\pi[e^{s|x|}] < \infty$ for some $s > 0$. We show that this is the case here.

From (7.3), for any $x_t \in \mathbf{X}$ we have that

$$\begin{aligned}
|x_{t+L\varepsilon} - x_t| &= |L\varepsilon^2 \nabla \log \pi(x_t)/2 + \varepsilon^2 \sum_{i=1}^{L-1} (L-i) \nabla \log \pi(x_{t+i\varepsilon}) + L\varepsilon p_t|, \\
&\leq L\varepsilon^2 |\nabla \log \pi(x_t)|/2 + \varepsilon^2 \sum_{i=1}^{L-1} (L-i) |\nabla \log \pi(x_{t+i\varepsilon})| + L\varepsilon |p_t|, \\
&\leq L\varepsilon^2 C/2 + \varepsilon^2 L(L-1)C/2 + L\varepsilon |p_t|.
\end{aligned}$$

Since $|p_t|$ is the norm of a Gaussian random variable whose variance does not depend on x_t, then Chebyshev's inequality gives the result. ■

In fact, this negative result can be extended to the idealised Hamiltonian Monte Carlo algorithm, where the true flow can be simulated exactly, as the following shows.

Proposition 7.4. *For a fixed integration time T, if $|\nabla \log \pi(x)| < C$ for every $x \in \mathbf{X}$, then the idealised Hamiltonian Monte Carlo method can produce a geometrically ergodic Markov chain* only *in the case where $\mathbb{E}_\pi[e^{s|x|}] < \infty$ for some $s > 0$.*

Proof: We can proceed as in the previous Proposition. Here, using Hamilton's equations, we have

$$x_{t+T} - x_t = \int_0^T p_{t+s} ds = \int_0^T \left[p_t + \int_0^s \nabla \log \pi(x_{t+u}) du \right] ds. \tag{7.5}$$

Taking the norm and using the upper bound gives

$$\begin{aligned}
|x_{t+T} - x_t| &\leq T|p_t| + \int_0^T \int_0^s |\nabla \log \pi(x_{t+u})| du ds, \\
&\leq T|p_t| + CT^2/2,
\end{aligned}$$

and again Chebyshev's inequality gives the result. ■

The above results apply to general target distributions of any dimension. However, from this point forward we restrict our attention to the one-dimensional exponential family introduced at the beginning of the chapter. We turn first to the special cases $\beta = 1$ and $\beta = 2$, corresponding to the

Laplace and Gaussian distributions. As a comment, we note that Hamilton's equations can in fact be integrated exactly in these scenarios, so the idealised algorithm can actually be employed. However, our interest is in methods which can be applied to a much broader class of targets, so we consider the leap-frog scheme here.

Theorem 7.5. *For the one-dimensional exponential family class of targets, the Hamiltonian Monte Carlo method produces a geometrically ergodic Markov chain in the case $\beta = 1$ (Laplace distribution), and provided a suitably small step-size ε is chosen, also in the case $\beta = 2$ (Gaussian distribution).*

In the Laplacian case, leapfrog dynamics actually exactly solve Hamilton's equations provided that the $x = 0$ boundary is not crossed. So in this case the method is simply a random walk with inwards drift, and the proposal is Gaussian, so it is straightforward to show that this is geometrically ergodic using the Lyapunov function $V(x) = e^{s|x|}$ for some $s > 0$. In the Gaussian case, the proposal is still Gaussian but does not exactly replicate Hamiltonian flow, so the acceptance probability must also be considered. However, in this case it can be shown that the algorithm still reduces to a *version* of the Metropolis-adjusted Langevin algorithm, and so will produce a geometrically ergodic chain provided care is taken with the tuning parameters.

In the case where $|\nabla \log \pi(x)|/|x| \to \infty$ as $|x| \to \infty$, then the Metropolis-adjusted Langevin algorithm fails to produce a geometrically ergodic chain, as proposals 'explode' in the tails, and as a result very few are accepted (see Section 4.2.2 for intuition and [109] for a proof). Intuitively this should also be the case in Hamiltonian Monte Carlo. The next result confirms this intuition.

Theorem 7.6. *For the one-dimensional exponential family class of targets, the Hamiltonian Monte Carlo method does not produce a geometrically ergodic Markov chain in the case $\beta > 2$.*

The basic intuition for the result is to show that provided $|x_t|$ is sufficiently large and $|p_t|$ small relative to it, then at the ith leapfrog step $|x_{t+i\varepsilon}| > 2|x_{t+(i-i)\varepsilon}|$ and $|p_{t+i\varepsilon}| > 2|p_{t+(i-1)\varepsilon}|$, meaning the probability of accepting a proposed move $(x_{t+L\varepsilon}, p_{t+L\varepsilon})$ can be made arbitrarily small by choosing $|x_t|$ large enough. To conclude the proof from here we simply note that as $|x_t|$ grows then the probability that $|p_t|$ will be small relative to it can be made arbitrarily large, as p_t is simply a Gaussian with fixed covariance and zero mean.

This result is not a property of the exact flow itself, which is in fact extremely efficient, but rather the numerical integrator. In this instance the period length gets smaller as x grows, meaning a smaller integration time is needed to reach the centre of the space. In this instance the leapfrog numerical scheme becomes unstable, resulting in a diverging numerical flow. The problem of numerically solving *stiff* systems such as this one is well-documented (e.g. [119])

The final result of this section concerns the remaining possible values for the parameter β in the class of targets under consideration.

Theorem 7.7. *For the one-dimensional exponential family class of targets, the Hamiltonian Monte Carlo method produces a geometrically ergodic Markov chain in the case* $1 < \beta < 2$.

In this case we use several concepts that were introduced in [109] to analyse the Metropolis-adjusted Langevin algorithm. If we write the HMC proposal here in the form $x_{t+L\varepsilon} = c_h(x_t, p_t) + L\varepsilon p_t$, and the corresponding MALA proposal in the form $y = c(x_t) + \varepsilon p_t$, then we show that $0 < c_h(x_t, p_t) < c(x_t)$ as $x_t \to \infty$ for almost all choices of p_t. This implies that typical HMC proposals will be closer to the centre of the space than those under a MALA scheme, and since the latter is geometrically ergodic then we can show from this that the former will be too. We also rely on the concept of *inwards convergence*, as in [109]. This restriction on the chain implies that as $|x_t| \to \infty$, all proposals that are closer to the centre of $\mathbf{X}$ will be accepted, whereas all that have a larger norm than $|x_t|$ could be rejected. We show in the proof that the HMC kernel satisfies this property here.

7.3 Changing integration times

The scheme we consider in this section is both *dynamic* and *idealised*. It is dynamic as the integration time changes based on the current position. It is idealised because we make two *unreasonable* assumptions. We first assume that we can solve Hamilton's equations exactly for the model in question. This can be relaxed with some additional work (in terms of choosing the correct acceptance rate for proposals, which will no longer be reversible), but is assumed for ease of exposition. Second we assume i) that any contour $C_{x_t,p_t} := \{(x,p) \in \mathbf{X} \times \mathbf{X} : H(x,p) = H(x_t,p_t)\}$ is a compact, disjoint union of simply connected components, that for a large enough x_t the contours will consist of a single connected component, and that the flow is periodic from any fixed starting point, and ii) that the *period length* ζ_{x_t,p_t} (the time taken to traverse the specific component of C_{x_t,p_t} in which (x_t,p_t)

lies) is known. Part i) is likely to be true for many statistical models of interest, but ii) will typically not be known outside of the case where $\pi(\cdot)$ is Gaussian (where $\zeta_{x_t,p_t} = \zeta$, as shown in a previous section). We discuss practical approaches to approximating ζ_{x_t,p_t} in Section 7.5.

At iteration i (with $x_t = x_{i-1}$), the *dynamic* Hamiltonian Monte Carlo implementation we consider consists of re-sampling $p_t \sim N(0,I)$, and then setting $x_i = \varphi^x_\tau(x_t,p_t)$, where $\tau \sim U[0,\zeta_{x_t,p_t}]$. In words, we flow along the Hamiltonian for τ units of time, where τ is a uniform random variable with maximum value the period length ζ_{x_t,p_t} (note that $\varphi_{\zeta_{x_t,p_t}}(x_t,p_t) = (x_t,p_t)$ here).

Firstly, note that μ^L-irreducibility is more straightforward to see here. To reach any set $A \in \mathscr{B}$ with $\mu^L(A) > 0$, we first consider the single contour C_{x_t,p_t}, and specifically the component of this contour that is connected to (x_t,p_t). Let C_{x_t} be the projection of this component onto $\mathbf{X}$. Then any nonempty set $A' \subset C_{x_t}$ has positive probability of occuring, as the next point is chosen from a density with support all of C_{x_t}. As the contours are eventually composed of single components, and cover the entire space, then for any A, the probability of choosing a contour for which this argument can be applied is greater than zero. Figure 7.2 offers more intuition.

To establish geometric ergodicity, we rely on conservation of the Hamiltonian, i.e.

$$\int U(x_{t+u},p_{t+u})\nu_{x_t,p_t}(du) + \int K(x_{t+u},p_{t+u})\nu_{x_t,p_t}(du) = \int H(x_{t+u},p_{t+u})\nu_{x_t,p_t}(du) = H(x_t,p_t). \tag{7.6}$$

Averaging over initial momentum choices for the case $K(p_t) = p_t^2/2$ gives

$$\int H(x_t,p_t)\mu^G(dp_t) = U(x_t) + 1/2.$$

We first introduce a result from the Physics literature which relates K and U. Using this we can relate the left hand side of (7.6) to

$$PU(x_t) = \int U(y)P(x_t,dy),$$

where P is the transition kernel under consideration. Since the right-hand side of (7.6) relates to the current value of $U(x_t)$, then our goal will be to construct a suitable Lyapunov function from the potential energy that will help us establish the necessary drift condition.

Theorem 7.8. *(Virial Theorem). Under Hamiltonian flow $(x_{t+s},p_{t+s}) = \varphi_s(x_t,p_t)$ we have*

$$\int x_{t+s}\frac{dU}{dx}(x_{t+s})\nu_{x_t,p_t}(ds) = 2\int K(p_{t+s})\nu_{x_t,p_t}(ds), \tag{7.7}$$

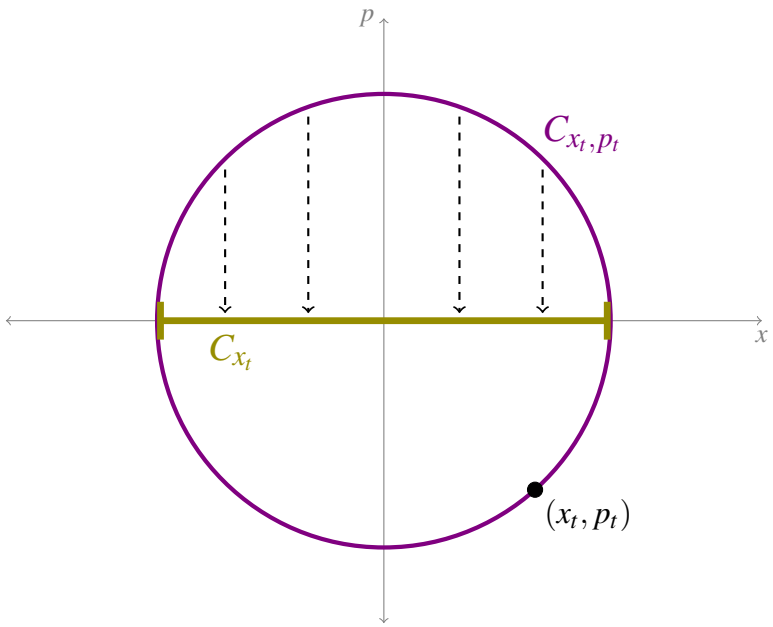

Figure 7.2: The contour $C_{x_t,p_t} = \{(y,z) \in \mathbb{R}^2 : y^2 + z^2 = 9\}$ for the Hamiltonian flow with Gaussian target $\pi(x) \propto e^{-x^2/2}$, with current point (x_t, p_t) lying on the disc of radius 3, and its projection onto the set $C_{x_t} = [-3,3]$.

where $\nu_{x_t,p_t}(ds) = \zeta_{x_t,p_t}^{-1} ds$ *denotes the Uniform distribution on* $[0, \zeta_{x_t,p_t}]$.

Proof: Define the *virial* function $G_t = x_t p_t$. From the fundamental theorem of Calculus we have

$$\int \dot{G}_{t+s} \nu_{x_t,p_t}(ds) = \zeta_{x_t,p_t}^{-1} \int_0^{\zeta_{x_t,p_t}} \dot{G}_{t+s} ds = \frac{G_{t+\zeta_{x_t,p_t}} - G_t}{\zeta_{x_t,p_t}} = 0.$$

In this case

$$\dot{G}_t = x_t \dot{p}_t + p_t \dot{x}_t = -x_t \frac{dU}{dx}(x_t) + p_t \frac{dK}{dp}(p_t),$$

meaning

$$\int x_{t+s} \frac{dU}{dx}(x_{t+s}) \nu_{x_t,p_t}(ds) = \int p_{t+s} \frac{dK}{dp}(p_{t+s}) \nu_{x_t,p_t}(ds).$$

Now simply note that

$$p_t \frac{dK}{dp}(p_t) = p_t^2 = 2K(p_t),$$

which, after substituting into the above equation, completes the proof. ■

Corollary 7.9. *For the one-dimensional exponential family of targets, and any* $(x_t, p_t) \in \mathbf{X} \times \mathbf{X}$*, we have*

$$2\int K(p_{t+s})\nu_{x_t,p_t}(ds) = \beta \int U(x_{t+s})\nu_{x_t,p_t}(ds). \tag{7.8}$$

Proof: Note that for this class of target distributions $U(x) = \beta^{-1}|x|^{\beta}$, so we have the relation

$$x_t \frac{dU}{dx}(x_t) = x_t \mathrm{sgn}(x_t)|x_t|^{\beta-1} = |x_t|^{\beta} = \beta U(x_t). \tag{7.9}$$

Substituting into (7.7) gives the result. ■

With these preliminaries, we can now state and prove the main result of this section.

Theorem 7.10. *For the one-dimensional exponential family class of targets, the dynamic Hamiltonian Monte Carlo method produces a geometrically ergodic Markov chain for any value of* $\beta > 0$*.*

Proof: Choose the Lyapunov function $V(x) = U(x) + 1$. Then

$$PV(x_t) = \int\int U(x_{t+s})\nu_{x_t,p_t}(ds)\mu^G(dp_t) + 1. \tag{7.10}$$

Note that by conservation of the Hamiltonian, we have

$$\int\int \left[U(x_{t+s}) + K(p_{t+s})\right]\nu_{x_t,p_t}(ds)\mu^G(dp_t) = \int H(x_t,p_t)\mu^G(dp_t) = U(x_t) + 1/2, \tag{7.11}$$

where we have used that $\int K(p_t)\mu^G(dp_t) = 1/2$. Using (7.8) and (7.10), the left-hand side of (7.11) can be written

$$\begin{aligned} PV(x_t) - 1 + \int\int K(p_{t+s})\nu_{x_t,p_t}(ds)\mu^G(dp_t) &= PV(x_t) - 1 + \beta(PV(x_t) - 1)/2, \\ &= (1+\beta/2)PV(x_t) - 1 - \beta/2. \end{aligned}$$

Substituting back into (7.11), simplifying, and dividing through by $(1+\beta/2)$ gives

$$PV(x_t) = \frac{1}{(1+\beta/2)}V(x_t) + \frac{1+\beta}{2+\beta}.$$

To complete the proof note that we can choose an $x \in \mathbf{X}$ such that for all $|y| \geq |x|$ we have

$$\frac{\beta/3}{1+\beta/2}V(y) > \frac{1+\beta}{2+\beta},$$

meaning

$$PV(x_t) \leq \left(\frac{1+\beta/3}{1+\beta/2}\right)V(x_t) + \left(\frac{1+\beta}{2+\beta}\right)\mathbb{1}_C(x_t),$$

with $C = (-x, x)$. ■

7.4 Proofs

The longer proofs of results stated in previous sections are given here, to aid readability of the main text. In each case we re-state the result and then provide a full proof.

7.4.1 Proof of Theorem 7.5

The Hamiltonian Monte Carlo method produces a geometrically ergodic Markov chain in the case $\beta = 1$ *(Laplace distribution), and provided a suitably small step-size* ε *is chosen, also in the case* $\beta = 2$ *(Gaussian distribution).*

Proof: In the $\beta = 1$ case we note that setting $T = L\varepsilon$, then the numerical map $\eta_{L\varepsilon}(x,p)$ and the exact flow $\varphi_T(x,p)$ are identical provided the flow does not cross the point $x = 0$. Noting that $\nabla \log \pi(x) = -\text{sgn}(x)$ here, then in the marginal case, for any fixed p_t there is a large enough x_t that, using (7.3) and (7.5) we have

$$\varphi_T^x(x_t, p_t) = x_t + T p_t - \int_0^T \int_0^s du ds = x_t + T p_t - T^2/2.$$

Setting $T = L\varepsilon$ gives

$$\varphi_T^x(x_t, p_t) = \eta_{L\varepsilon}^x(x_t, p_t).$$

Because of this, it is straightforward to see that here the Hamiltonian $H(x,p) = |x| + p^2/2 + const.$ is preserved exactly by the numerical flow induced from leapfrog dynamics, meaning that the acceptance rate is 1 here when the above equation is satisfied. So using the Lyapunov function $V(x) = e^{s|x|}$ for some $s > 0$ gives

$$\begin{aligned}\limsup_{x\to\infty} \frac{PV(x)}{V(x)} &= \limsup_{x\to\infty} \int e^{s(|x-(L\varepsilon)^2/2+p|-|x|)} \mu^G(dp), \\ &\leq e^{-s(L\varepsilon)^2/2} \limsup_{x\to\infty} \int e^{s|p|} \mu^G(dp),\end{aligned}$$

where $\mu^G(\cdot)$ is a standard Gaussian measure. Using properties of truncated Gaussian distributions (see Appendix D) gives for large enough x and small enough $s > 0$

$$\frac{PV(x)}{V(x)} \leq 2e^{s(s-1)(L\varepsilon)^2/2} \Phi(sL\varepsilon) < 1,$$

as required. A similar argument can be made as $x \to -\infty$.

In the Gaussian case we have $\nabla \log \pi(x) = -x$. In the paper [6], the authors note that here the transition kernel can actually be written

$$\begin{aligned} x_{t+L\varepsilon} &= \cos(\theta L) x_t + \frac{\sin(\theta L)}{\sqrt{1-\varepsilon^2/4}} p_t, \\ p_{t+L\varepsilon} &= -\sqrt{1-\varepsilon^2/4}\sin(\theta L) x_t + \cos(\theta L) p_t, \end{aligned} \tag{7.12}$$

where $\theta = \arccos(1-\varepsilon^2/2)$. Note that from the first equation

$$-p_t^2/2 = -\frac{1-\varepsilon^2/4}{2\sin^2(\theta L)} \left(x_{t+L\varepsilon} - \cos(\theta L) x_t\right)^2 = \log q(x_t | x_{t+L\varepsilon}).$$

Similarly, owing to the reversibility of the leapfrog flow, we have

$$\log q(x_{t+L\varepsilon} | x_t) = -p_{t+L\varepsilon}^2/2,$$

meaning

$$q(x_t | x_{t+L\varepsilon}) / q(x_{t+L\varepsilon} | x_t) = \exp\left(-\frac{1}{2}\left[p_{t+L\varepsilon}^2 - p_t^2\right]\right).$$

This explicitly shows that the Hamiltonian Monte Carlo method can simply be thought of as a regular Metropolis–Hastings method with proposal (7.12) here. In this case it can actually be seen as a MALA proposal, which by Theorem 4.1 in [109] is geometrically ergodic provided $|\cos(\theta L)| < 1$.

■

7.4.2 Proof of Theorem 7.6

The Hamiltonian Monte Carlo method does not produce a geometrically ergodic Markov chain in the case $\beta > 2$.

Proof: We first show that the leapfrog integrator pushes proposals further out into the tails at an increasing rate as $|x_t|$ grows, and then that this implies that the rejection probability $r(x_t) \to 1$ as $x_t \to \infty$.

Suppose that

$$\varepsilon p_t / x_t < 3/2 \quad \text{and} \quad \varepsilon^2 |x_t|^{\beta-2} > 9.$$

Then after a single leapfrog step, $(x_{t+\varepsilon}, p_{t+\varepsilon}) = \eta_\varepsilon(x_t, p_t)$, we have that

(a) $x_{t+\varepsilon} < -2x_t$,
(b) $\varepsilon p_{t+\varepsilon} / x_{t+\varepsilon} < 3/2$,
(c) $|p_{t+\varepsilon}| > 2|p_t|$.

To see (a), note that the position update here is

$$x_{t+\varepsilon} = x_t \left(1 - \varepsilon^2 |x_t|^{\beta-2}/2 + \varepsilon p_t / x_t\right) \leq x_t \left(1 - 9/2 + 3/2\right) = -2x_t$$

Note that this also implies that $-x_t/x_{t+\varepsilon} < 1/2$. As the integrator is reversible, we can also write

$$x_t = x_{t+\varepsilon} - \varepsilon^2 x_{t+\varepsilon} |x_{t+\varepsilon}|^{\beta-2}/2 - \varepsilon p_{t+\varepsilon}.$$

Rearranging and dividing by $x_{t+\varepsilon}$ gives

$$\varepsilon p_{t+\varepsilon}/x_{t+\varepsilon} = 1 - \varepsilon^2 |x_{t+\varepsilon}|^{\beta-2}/2 - x_t/x_{t+\varepsilon} < 3/2,$$

which establishes (b).

To see (c), note from the marginal momentum update (7.2) that

$$\frac{p_{t+\varepsilon/2}}{x_t} = \frac{p_t}{x_t} - \frac{\varepsilon}{2}|x_t|^{\beta-2} < \frac{p_t}{x_t}, \tag{7.13}$$

and since

$$x_{t+\varepsilon} - x_t = \varepsilon \left(p_t + \varepsilon \nabla \log \pi(x_t)/2\right) = \varepsilon p_{t+\varepsilon/2},$$

then rearranging gives

$$\frac{p_{t+\varepsilon/2}}{x_t} = \frac{1}{\varepsilon}\left(\frac{x_{t+\varepsilon}}{x_t} - 1\right) < -3/\varepsilon, \tag{7.14}$$

meaning $|p_{t+\varepsilon}/x_t| > 3/\varepsilon$. Combining with (7.13) and using the condition $\varepsilon p_t/x_t < 3/2$ gives

$$\left|\frac{p_t}{x_t}\right| < \left|\frac{p_{t+\varepsilon/2}}{x_t}\right|.$$

Finally, from the final leapfrog step for momentum $p_{t+\varepsilon}$ we have that

$$\begin{aligned}
\frac{p_{t+\varepsilon}}{x_t} &= \frac{p_{t+\varepsilon/2}}{x_t} - \frac{\varepsilon}{2}\frac{x_{t+\varepsilon}}{x_t}|x_{t+\varepsilon}|^{\beta-2} \\
&= \frac{p_{t+\varepsilon/2}}{x_t} - \frac{\varepsilon}{2}\left(1 + \varepsilon\frac{p_{t+\varepsilon/2}}{x_t}\right)|x_{t+\varepsilon}|^{\beta-2} \\
&= \frac{p_{t+\varepsilon/2}}{x_t} - \frac{\varepsilon}{2}\left(1 + \frac{\varepsilon}{2}\frac{p_{t+\varepsilon/2}}{x_t}\right)|x_{t+\varepsilon}|^{\beta-2} - \frac{\varepsilon^2}{4}\frac{p_{t+\varepsilon/2}}{x_t}|x_{t+\varepsilon}|^{\beta-2}.
\end{aligned}$$

Using (7.14) we have that

$$-\frac{\varepsilon^2}{4}\frac{p_{t+\varepsilon/2}}{x_t}|x_{t+\varepsilon}|^{\beta-2} - \frac{\varepsilon}{2}|x_{t+\varepsilon}|^{\beta-2} > \frac{\varepsilon}{4}|x_{t+\varepsilon}|^{\beta-2} > 0,$$

meaning

$$\frac{p_{t+\varepsilon}}{x_t} > \frac{p_{t+\varepsilon/2}}{x_t}\left(1 - \frac{\varepsilon^2}{2}|x_{t+\varepsilon}|^{\beta-2}\right) > -2\frac{p_{t+\varepsilon/2}}{x_t},$$

which establishes (c).

By induction, we can see further that under these conditions, after L leapfrog steps,

$$|x_{t+L\varepsilon}| > 2^L |x_t| \quad \text{and} \quad |p_{t+L\varepsilon}| > 2^L |p_t|.$$

In other words, for any such (x_t, p_t), the resulting leapfrog trajectories will rapidly get larger in magnitude, diverging from the true oscillating trajectories.

We now show that this exploding in magnitude implies $r(x_t) \to 1$ as $x_t \to \infty$. Writing $z_t = (x_t, p_t)$, recall that the probability of accepting a proposed move is

$$\alpha(z_t, z_{t+L\varepsilon}) = 1 \wedge \exp\left(\beta^{-1}|x_t|^\beta - \beta^{-1}|x_{t+L\varepsilon}|^\beta + |p_t|^2/2 - |p_{t+L\varepsilon}|^2/2\right).$$

If $\varepsilon p_t / x_t < 3/2$ then provided $\varepsilon^2 |x_t|^{\beta-2} > 9$ we have

$$\begin{aligned}
\alpha(z_t, z_{t+L\varepsilon}) &< \exp\left(\beta^{-1}(1-2^L)|x_t|^\beta + (1-2^L)|p_t|^2/2\right) \\
&\leq \exp\left(-\beta^{-1}|x_t|^\beta - |p_t|^2/2\right), \\
&\leq \exp\left(-\beta^{-1}|x_t|^\beta\right)
\end{aligned}$$

where we have used the fact that $L \geq 1$. This quantity clearly tends to zero as $x_t \to \infty$. So to complete the proof we simply note that

$$\mathbb{P}[\varepsilon p_t / x_t < 3/2] = \mathbb{P}[p_t < 3x_t/2\varepsilon] = \Phi(3x_t/2\varepsilon) \to 1$$

as $x_t \to \infty$, where Φ is the standard Gaussian cumulative distribution function. ■

7.4.3 Proof of Theorem 7.7

The Hamiltonian Monte Carlo method produces a geometrically ergodic Markov chain in the case $1 < \beta < 2$.

Proof: We first recall the proof of geometric ergodicity of MALA in this scenario, and then extend this proof to the HMC case.

MALA Proof.

Notation: Define $A(x) = \{y \in \mathbf{X} : \alpha(x,y) = 1\}$, $R(x) = A(x)^c$, $I(x) = \{y \in \mathbf{X} : |y| \leq |x|\}$, and $c(x) = x + h\nabla \log \pi(x)/2$ as the mean next candidate step. We say $A(x)$ 'converges inwards' if

$$\lim_{|x|\to\infty} \int_{A(x)\triangle I(x)} Q(x, dy) = 0,$$

where $A(x) \triangle I(x) = (A(x) \cap I(x)^c) \cup (R(x) \cap I(x))$ is the *symmetric difference* of $A(x)$ and $I(x)$. this implies that in the limit all inwards proposals are accepted and all outwards proposals might be rejected.

Setting $V(x) = e^{s|x|}$ for some $s > 0$, we have

$$\begin{aligned}\frac{PV(x)}{V(x)} &= \int_{A(x)} e^{s(|y|-|x|)} Q(x,dy) + \int_{R(x)} e^{s(|y|-|x|)} \alpha(x,y) Q(x,dy) + \int_{R(x)} (1-\alpha(x,y)) Q(x,dy), \\ &= \int_{\mathbb{R}} e^{s(|y|-|x|)} Q(x,dy) + \int_{R(x)} \left[1 - e^{s(|y|-|x|)}\right] (1-\alpha(x,y)) Q(x,dy), \\ &\leq \int_{\mathbb{R}} e^{s(|y|-|x|)} Q(x,dy) + \int_{R(x)\cap I(x)} Q(x,dy).\end{aligned}$$

The first term asymptotes to $e^{s(|c(x)|-|x|)+s^2/2h}$ which is certainly less than one whenever $|c(x)| - |x| \to -\infty$ as $|x| \to \infty$. The second term disappears under the assumption that $A(x)$ converges inwards. It is shown in [109] that both of these conditions hold for the class of target distributions under consideration here. ■

Extension to HMC.

Extra notation: Define $c_h(x_t, p_t) := x_t + L\varepsilon^2 \nabla \log \pi(x_t)/2 + \varepsilon^2 \sum_{i=0}^{L-1} (L-i) \nabla \log \pi(x_{t+i\varepsilon})$, so that $x_{t+L\varepsilon} = c_h(x_t, p_t) + L\varepsilon p_t$. Also set $\psi(x_t, p_t) := L\varepsilon^2 \nabla \log \pi(x_t)/2 + \varepsilon^2 \sum_{i=1}^{L-1} (L-i) \nabla \log \pi(x_{t+i\varepsilon})$ for the increment term. We also define the set $B(x) := (-|x|^\delta, |x|^\delta)$ for some $1 > \delta > \max(\beta - 1, 1/2)$, and let $\mu^G(\cdot)$ denote a standard Gaussian measure.

We extend the MALA result in 4 steps. Starting from the expression

$$\frac{PV(x)}{V(x)} \leq \int_{\mathbb{R}} e^{s(|y|-|x|)} Q(x,dy) + \int_{R(x)\cap I(x)} Q(x,dy), \tag{7.15}$$

we do the following:

1. Establish that for $p_t \in B(x_t)$, and large enough x_t we have $0 \leq c_h(x_t, p_t) \leq c(x_t) \leq x_t$,
2. Use this to bound the first integral in (7.15) with that for MALA for $p_t \in B(x_t)$
3. Show that the integral is negligible outside this region
4. Establish inwards convergence, meaning the last integral in (7.15) is also negligible.

1. Noting that $\nabla \log \pi(x) = -\text{sgn}(x_t)|x_t|^{\beta-1}$, and taking $x_t \gg 0$, we have

MALA: $x' = c(x_t) + hp$.

HMC: $x_{t+L\varepsilon} = c_h(x_t, p_t) + L\varepsilon p_t$.

Setting $h = \varepsilon\sqrt{L}$ gives

$$c_h(x_t, p_t) = c(x_t) - \varepsilon^2 \sum_{i=1}^{L-1} (L-i)\text{sgn}(x_{t+i\varepsilon})|x_{t+i\varepsilon}|^{\beta-1}.$$

To establish that $0 \leq c_h(x_t, p_t) \leq c(x_t)$ for $p_t \in B(x_t)$, it is therefore sufficient to show the following for $i \in \{1, ..., L-1\}$

(i). $\text{sgn}(x_{t+i\varepsilon}) = 1$, implying or equivalently each $x_{t-i\varepsilon} > 0$
(ii). $\varepsilon^2 \sum_{i=1}^{L-1} (L-i)\text{sgn}(x_{t+i\varepsilon})|x_{t+i\varepsilon}|^{\beta-1} < c(x_t)$.

We show both of these by establishing upper and lower bounds for $x_{t+i\varepsilon}$ and in the limit for large x_t. In what follows we use the symbol $\gtrsim$ to mean 'asymptotically greater than or equal to', so that $f(x_t) \gtrsim x_t$ means that for all x_t sufficiently large $f(x_t) \geq x_t$. We also define $\lesssim$ analogously. We also let $\{\lambda_1, ..., \lambda_{L-1}\}$ and $\{\rho_1, ..., \rho_{L-1}\}$ be two sequences of constants that satisfy $1 > \lambda_{L-1} > ... > \lambda_1 > 0$ and $1 > \rho_{L-1} > ... > \rho_1 > 0$.

After a single leapfrog step we have

$$x_{t+\varepsilon} = x_t - \varepsilon^2 x_t^{\beta-1}/2 + \varepsilon p_t.$$

After noting that $p_t/x_t \to 0$ as $x_t \to \infty$, it is straightforward to see that here

$$(1-\lambda_1)x_t \lesssim x_{t+\varepsilon} \lesssim (1+\lambda_1)x_t.$$

For the momentum we have

$$p_{t+\varepsilon} = p_t - \varepsilon x_t^{\beta-1}/2 - \varepsilon x_{t+\varepsilon}^{\beta-1}/2.$$

Noting that $1 > \delta > \beta - 1$, then we have

$$-(1+\rho_1)x_t^{\delta} \lesssim p_{t+\varepsilon} < p_t.$$

Continuing the argument, after the next leapfrog step we have

$$x_{t+2\varepsilon} = x_{t+\varepsilon} - \varepsilon^2 x_{t+\varepsilon}^{\beta-1}/2 + \varepsilon p_{t+\varepsilon}.$$

Using the same arguments and the bounds on $p_{t+\varepsilon}$ gives

$$(1-\lambda_2)x_t \lesssim x_{t+2\varepsilon} \lesssim (1+\lambda_2)x_t$$

and similarly for $p_{t+2\varepsilon}$ we have

$$-(1+\rho_2)x_t^{\delta} \lesssim p_{t+2\varepsilon} < p_t.$$

By induction, we have

$$(1-\lambda_{L-1})x_t \lesssim x_{t+(L-1)\varepsilon} \lesssim (1+\lambda_{L-1})x_t.$$

Since every term $x_{t+i\varepsilon}$ is positive, this establishes (i). As they are each of the same order as x_t, then (ii) is also true for large enough x_t.

2. The integral in question can be written

$$\int_{B(x_t)} e^{s(|y(p)|-|x_t|)}\mu^G(dp) + \int_{B(x_t)^c} e^{s(|y(p)|-|x_t|)}\mu^G(dp).$$

In $B(x_t)$ we have

$$|y(p)| = |c_h(x_t,p_t)+L\varepsilon p_t| \le |c_h(x_t,p_t)|+L\varepsilon|p_t| \le |c(x_t)|+L\varepsilon|p_t|,$$

meaning

$$\int_{B(x_t)} e^{s(|y(p)|-|x_t|)}\mu^G(dp_t) \le e^{s(|c(x)|-|x_t|)}\int_{B(x_t)} e^{sL\varepsilon p_t}\mu^G(dp_t) \le e^{s(|c(x)|-|x_t|)}\left[2e^{(sL\varepsilon)^2/2}\Phi(sL\varepsilon)\right],$$

which can be made arbitrarily small as $|c(x)|-|x_t| \to -\infty$ as $|x_t| \to \infty$.

3. We can actually extend and simplify the above argument for the purposes of this step. Clearly we have $|x_{t+\varepsilon}| = \Theta(\max(|x_t|,|p_t|))$ and $|p_{t+\varepsilon}| = \Theta(\max(|p_t|,|x_t|^{\beta-1}))$. It follows that $|x_{t+L\varepsilon}| = \Theta(\max(|x_t|,|p_t|)$. This means that for some $C<\infty$ for $|p_t| \ge |x_t|^\delta$ we have

$$\begin{aligned}
\exp\left(s|x_{t+L\varepsilon}|-s|x_t|-\frac{1}{2}|p_t|^2\right) &\lesssim \exp\left(C\max(|x_t|,|p_t|)-\frac{1}{2}|p_t|^2\right),\\
&= \exp\left(|p_t|\left(C\frac{\max(|x_t|,|p_t|)}{|p_t|}-\frac{1}{2}|p_t|\right)\right),\\
&\lesssim \exp\left(|p_t|\left(C|x_t|^{1-\delta}-\frac{1}{2}|p_t|\right)\right)
\end{aligned}$$

Provided $\delta > 1/2$, then for suitably large $|x_t|$ we have $|x_t|^{1-\delta}-|x_t|^\delta/2 < -1$, so we can write

$$\exp\left(|p_t|\left(C|x_t|^{1-\delta}-\frac{1}{2}|p_t|\right)\right) \lesssim \exp\left(-|p_t|\right),$$

meaning

$$\int_{B(x_t)^c} e^{s(|x_{t+L\varepsilon}|-|x_t|)}\mu_x(dp_t) \lesssim \int_{B(x_t)^c} e^{-|p_t|}dp_t = 2e^{-|x_t|^\delta},$$

which becomes negligibly small as $|x_t| \to \infty$, as required.

4. We call the acceptance *ratio* for a proposal

$$\exp\left(\beta^{-1}\left[|x_t|^\beta - |c_h(x_t,p_t)+L\varepsilon p_t|^\beta\right]+\frac{1}{2}p_t^2-\frac{1}{2}p_{t+L\varepsilon}^2\right).$$

This is simply the acceptance probability without the minimum term. For $p_t \in B(x_t)$ it is shown in (1) that the absolute value symbols here are not needed. As $\beta > 1$ then for $|x_t| > |x_{t+L\varepsilon}|$ here then we have

$$x_t^\beta - (c_h(x_t, p_t) + L\varepsilon p_t)^\beta > (x_t^\beta - x_t) - \psi(x_t, p_t) - L\varepsilon p_t,$$

meaning the acceptance ratio in this region can be lower bounded by

$$\exp\left(\beta^{-1}\left[(x_t^\beta - x_t) - \psi(x_t, p_t) - L\varepsilon p_t\right] + \frac{1}{2}p_t^2 - \frac{1}{2}p_{t+L\varepsilon}^2\right)$$

Using the order arguments from step (2) means that dividing through by $|x_t|^\beta$ gives

$$\exp\left(|x_t|^\beta \beta^{-1}\left[1 - \delta_1\right] + \delta_2\right),$$

where the constants δ_1 and δ_2 can be made arbitrarily small by simply choosing x_t large enough. In the limit this is greater than one here meaning all proposals will be accepted.

Since $\int_{B(x_t)^c} Q(x, dy) \to 0$ as $x_t \to \infty$, then we have the established result, meaning the last integral in (7.15) is negligible. An analogous argument can be constructed in the case $x_t \to -\infty$.

■

7.5 Discussion & Extensions

We have established some rigorous guarantees for the Hamiltonian Monte Carlo method here, which have in turn suggested new directions in the design of Hamiltonian Monte Carlo algorithms and appropriate rules for setting tuning parameters. The Probabilist and Statistician Persi Diaconis has recently suggested that such analysis would be of great use to the field [29, 31], in one case exclaiming 'Someone should take up this challenge!' [29]. We are unaware of many similar examples of analysing the method, the works [30, 50, 19] being notable exceptions, which focus either on the more foundational properties of ergodicity in general or on specific statistical models or tail behaviours. An ultimate goal here is of course to establish sharp nonasymptotic bounds, and hence answer the all important question 'for how long should I run my algorithm?' We hope that the present contribution is both a useful and nontrivial step in this direction.

Below we discuss further work which would compliment the results of this chapter, referring to each case separately.

7.5.1 Static case

In all cases except target distributions which are not log-concave in the tails, the ergodicity results presented here are restricted to the one-dimensional exponential family class of targets. With some moderate effort, however, the results can be extended. To generalise to higher dimensions is actually much more straightforward for gradient-based proposals than those based on random walks, owing to the convenience of working with norms, and the general criteria introduced in [109]. The $1 < \beta < 2$ result should be extendable to any target distribution for which $|c_h(x_t, p_t)|$ is nearly always smaller than $|c(x_t)|$ in the tails, provided that the *inwards convergence* property holds. With this condition it can be shown that the drift condition for Hamiltonian Monte Carlo should hold whenever a similar condition holds for the Metropolis-adjusted Langevin algorithm. Similarly the $\beta > 2$ case should be generalisable to any distribution for which $|\nabla \log \pi(x)|$ grows at a faster than linear rate in $|x|$, as in the MALA case. Both of these extensions are immediate goals.

7.5.2 Dynamic case

The dynamic results are striking, and there are several potential avenues of further research here. The unrealistic assumption that the period lengths are known can be relaxed in the scenario that an 'approximate' form of the Virial theorem holds, i.e. the case

$$\int \dot{G}_{t+u} \nu_{x_t, p_t}(du) < \delta,$$

for some $\delta > 0$, rather than being exactly equal to zero. This approach enables practical schemes to be designed in which Hamilton's equations are integrated numerically until the criterion is satisfied, and then the next candidate position in the chain is sampled from within this trajectory. The above equation can also be satisfied in cases where the Hamiltonian flow may not actually be periodic, which strengthens the ergodicity results here, but also raises challenges with regards to irreducibility, which may not be straightforward to establish.

Of course, the dynamic scheme suggested here is not the only way in which the integration time can be increased when $|x_t|$ grows. Other more straightforward approaches could be to make the number of leapfrog steps satisfy

$$L(x_t) \propto |\nabla \log \pi(x_t)|^{-1} |x_t|,$$

in some appropriate way. To ensure that the resulting Markov chain targets the correct invariant

distribution, the ratio $q(x|y)/q(y|x)$ would also need to be computed when evaluating proposals here, so designing a scheme in which this evaluation is straightforward is a further consideration.

Another popular implementation of the Hamiltonian Monte Carlo algorithm is the No-U-turn sampler [49], which is used in the Stan software [123], and in which the integration time is also dynamically chosen. Empirical evidence given in [49] suggests that the method is very efficient, and it may be that the pragmatic criterion of integrating forward until a 'U-turn' is made actually corresponds to increasing the integration time until geometric ergodicity is satisfied in many cases (though unlikely in general due to pathological examples with contours of high curvature). If this were the case then the theory established in this chapter would give some rigorous justification for the impressive performance of the method, so we aim to explore whether this is the case with some simple examples as further work.

7.5.3 Stiff bounds and uses for practitioners

The light-tailed case in which $\beta > 2$ is a challenge for ordinary gradient-based samplers, and Hamiltonian Monte Carlo appears to be no exception. However, for the Metropolis-adjusted Langevin algorithm some nonasymptotic guarantees hold even in this instance. It is shown in [15] that for any initial position x, a step-size h can be chosen to be small enough that the MALA chain will explore the ball centred at 0 with radius $|x|$ at a geometric rate. This gives some guarantees in this case, with the obvious intuition that for any fixed x a step-size h can be chosen such that $|h^2 \nabla \log \pi(x)/2|$ is of a similar size to x, making the MALA proposal a sensible one. The downside is that in practice for small choices of h exploration will be very slow in the centre of the space, meaning the geometric rate r will be close to one.

It seems that with some work these results could be extended to the case of Hamiltonian Monte Carlo, and that the rewards would potentially be greater here. The same intuition applies as in the MALA case, but this time the step-size parameter is ε. So there is an option to find a small enough ε, and then choose a number of leapfrog steps L such that the resulting integration time $L\varepsilon$ is still long enough for fast exploration of the state space. Of course the additional costs in the case ε is small may still be prohibitive. The guarantees provided by such an approach are still also problem specific and rely on choosing expectations for which the ball of radius $|x|$ is an appropriate substitute for the state space, which is a very difficult condition to check in practice. Nonetheless, such a result would still likely improve understanding further of the Hamiltonian Monte Carlo method.

Many of the bounds constructed here rely on showing that HMC is 'strictly more efficient' than MALA, in some sense. Such a direct comparison would be arguably of the most use to practitioners when choosing which method to use on a given problem. We have not categorically established cases here in which the Hamiltonian Monte Carlo method will estimate any expectation of interest more efficiently than the Metropolis-adjusted Langevin, as the bounds we have rely on certain choices of Lyapunov function, however such a comparison would seem possible in principle by directly considering the return times of each method to some typical set, rather than bounding these through specific choices of Lyapunov function.

Chapter 8

Summary and future directions

Here we provide a short summary of the contributions in each of Chapters 5, 6 and 7, along with possible avenues for further research.

8.1 Langevin diffusions

In Chapter 5 we explore Langevin diffusions for Monte Carlo sampling.

8.1.1 Contributions

(I) We highlight that the Langevin diffusion with position-dependent volatility reported in [108] and [43] does not in fact have the desired limiting distribution in general, and derive a simpler diffusion which does, given by

$$dX_t = \frac{1}{2}G^{-1}(X_t)\nabla \log \pi(X_t)dt + \Lambda(X_t)dt + \sqrt{G^{-1}(X_t)}dB_t, \text{ where } \Lambda_i(X_t) = \frac{1}{2}\sum_{j=1}^{n} \partial_j G^{-1}_{ij}(X_t),$$

using the convention $\partial_j := \partial/\partial x_j$. We derive this result using techniques from stochastic analysis.

(II) We then derive the same diffusion using the techniques of Riemannian geometry, showing that it can also be viewed as the image of $dX_t = \nabla \log \pi(X_t)dt + \sqrt{2}dB_t$ when mapped onto a Riemannian manifold with metric $G(x)$. In doing this we are able to pinpoint where errors

were made in previous derivations of this object in [108] and [43].

(III) Finally we consider three popular choices of $G(x)$ in the literature: the *negative Hessian*, the *truncating* metric and the *linearising* metric. We discuss how each of these choices change the ergodic properties of the resulting Langevin diffusion. We provide both intuitive and rigorous results, in both the simple one-dimensional class of *Exponential family* distributions and a two-dimensional case which contains features which are common of hierarchical models.

8.1.2 Future directions

It is now known that a judicious choice of $G(x)$ can lead to Markov chains which are geometrically ergodic for a wide variety of tail behaviours [62]. An obvious mathematical question that remains is which choice is 'optimal' in a given scenario. This is likely to depend on both the function being estimated and the definition of optimality chosen. On a general note, a more direct comparison of the 'speed' of different Langevin diffusions than that provided here would likely give some useful insights.

A related but more computational question involves whether choice of $G^{-1}(x)$ which are *sparse*, in the sense that resulting matrix operations are not cubic in the dimension n of the state space, can still produce fast converging Markov chains, and indeed in which scenarios this is the case. It is shown here that different choices influence both the magnitude and direction of the 'drift' of the resulting Markov chain, and to what extent these two features can be adapted without a cubic implementation cost in dimension is a relevant open question.

Langevin diffusions can be used as both objects on which to based MCMC methods, or as the limiting process for an MCMC method. It has recently been shown in [7] that a Langevin diffusion with position-dependent volatility is the limiting process of the Random Walk Metropolis on certain classes of target distributions. A better understanding of when such limits and how this translates to choosing optimal parameters in an algorithm could lead to useful insights for practitioners.

8.2 Random Walk Metropolis with position-dependent proposal covariance

In Chapter 6 we consider a Metropolis–Hastings method with proposal $x' \sim N(x, hG^{-1}(x))$, which a variant of the Random Walk Metropolis in which the proposal variance is allowed to change with position. We discuss how allowing the covariance to change with position can change the conditions on $\pi(x)$ under which the resulting Markov chain will converge geometrically quickly to its limiting distribution, either for better or worse.

8.2.1 Contributions

(I) In one dimension we establish that if the variance is allowed to grow at $\Theta(|x|^\gamma)$ for some $0 < \gamma < 2$ then the method will produce a geometrically ergodic Markov chain provided the tails of the distribution of interest $\pi(\cdot)$ are such that $\pi(x) \leq \exp\left(-|x|^\beta\right)$ for some $\beta > 1 - \gamma/2$ for all $|x| \geq L$ for some large $L > 0$. If $\gamma = 2$ then we show that provided a small enough choice of the step-size h is made, the method will produce a geometrically ergodic chain provided that $\pi(x) \leq |x|^{-p}$ for all $|x| \geq L$, for some $p > 1$.

(II) We also establish some negative results: if the variance grows at a faster rate than $C|x|^2$ for some $C < \infty$ then we show that the algorithm can never produce a geometrically ergodic chain. We also show that in any dimension then a necessary condition for geometric ergodicity is that $\varepsilon^{s|x|}\pi(dx) < \infty$ for some $s > 0$ in the case where each element of then this will also be true if each element of $G^{-1}(x)$ is bounded above.

(II) Lastly we construct a simple two-dimensional density in which probability concentrates on an ever narrower 'ridge' as $|x|$ increases (a known scenario in which the ordinary Random Walk Metropolis can perform poorly). We show that a uniform proposal on a spherical disc will not produce a geometric converging chain, while for a uniform proposal over an elliptical disc, with the shape of ellipse dependent on the current position, the opposite is true. The first method is a version of the ordinary Random Walk Metropolis, while the second resembles the variant with changing covariance.

8.2.2 Future directions

General results in dimensions greater than one would of course be the natural next step in this work. In particular it would be interesting to understand whether a lack of smoothness of the contours of $\pi(x)$ can be corrected for 'arbitrarily well' by exploiting the position-dependent covariance framework, as this is a known failing of the Random Walk Metropolis and is a common feature of hierarchical models. This would also inform what an appropriate choice of $G^{-1}(x)$ in general should be. Of course a choice which involved as little information about $\pi(x)$ as possible is preferable, so that the method can be used in as wide a variety of scenarios as possible.

There is some intuition for believing that a judicious choice of $G^{-1}(x)$ might also produce a method that scales more favourably with dimension than the ordinary Random Walk Metropolis. A choice of covariance matrix for which the principal eigenvector $v(x)$ is asymptotically parallel to x and for which all other eigenvalues shrink to zero as $|x|$ grows should result in a method which proposes moves 'in the right direction' fifty percent of the time when in the tails. In large dimensions there are many erroneous directions in which random walk proposals can be, and the idea is that a changing $G^{-1}(x)$ can alleviate this to some extent. Making such an argument rigorous would be an interesting challenge.

The analysis here has shown that adaptive methods in which $G^{-1}(x)$ is learned based on computing a weighted empirical covariance of past samples and in which the adaptation stops after a fixed period of time are unlikely to result in different ergodicity properties to proposals with a fixed covariance. The reason is that one can always move far enough into the tails of the distribution that no past samples exist in this region. Ergodic rates, which inherently depend on the tails, are perhaps therefore not the most appropriate tools for analysing whether there is a benefit to such approaches. Two potential avenues for future research here are (i) considering whether allowing infinite adaptation on a compact state space, as in [46], can lead to more favourable ergodic properties, and (ii) using some other tools, such as expected squared jump distance analysis or diffusion limits, to compare relative efficiency of Markov chain exploration in the centre of the space.

8.3 Stability of Hamiltonian Monte Carlo

In Chapter 7 we consider the Hamiltonian Monte Carlo method, and establish conditions under which stochastic stability properties such as π-irreducibility and geometric ergodicity will and will not hold for certain classes of models.

8.3.1 Contributions

(I) By writing the HMC transition on the *marginal* position space, we show that under suitable assumptions on $\pi(\cdot)$, provided the gradient term $\|\nabla \log \pi(x)\|$ is continuous and grows at most linearly (with suitable integration time chosen in the linear case) then the method with a fixed integration time will produce a π-irreducible Markov chain, and all compact sets will be small.

(II) For geometric ergodicity, we firstly establish that for any model in which $\|\nabla \log \pi(x)\| \leq M$ for any $x \in \mathbf{X}$ then this can only happen in the case where $\int e^{s|x|}\pi(dx) < \infty$ for some $s > 0$. We also show that this is true if a perfect integrator were available. We then consider the one-dimensional exponential family class of targets

$$\pi(x) \propto \exp\left(-\beta^{-1}|x|^{\beta}\right), \quad \beta > 0.$$

For this class we show that HMC with a fixed integration time will produce a geometrically ergodic Markov chain if $1 \leq \beta \leq 2$, and will not otherwise.

(III) We then consider dynamic integration times, which are used in various implementations of HMC such as the No-U-Turn Sampler [49]. In an idealised scenario, we show that if the integration time is allowed to increase at a suitable rate and a perfect integrator is available then a geometrically ergodic chain can be constructed for any value $\beta > 0$.

8.3.2 Future directions

The immediate follow-on to this work is to extend the fixed integration time results to more general targets of arbitrary dimension. At the time of writing the thesis chapter this had not been completed, however it has now and as of January 2016 a preprint is in preparation. In essence the conditions are that $|\nabla \log \pi(x)|$ grows at most linearly (with careful choice of integration time in the linear case),

that the gradient points towards the centre of the space when $|x|$ is large, and that the algorithm 'converges inwards' in the sense of [109].

The dynamic integration time results are intriguing, but at present do not take into account the potential extra computational cost. To increase the integration time $T = L\varepsilon$ in the leapfrog method, one can either increase ε at no extra cost but reduced accuracy, or increase L at extra computational cost. The right choice is not obvious. It is also not clear how such an algorithm compares with simply considering more than one transition of a method with a fixed integration time. If the dynamic method take $2L\varepsilon$ steps in a certain region of the space, it is not clear how this compares to two transitions with a step-size of $L\varepsilon$, or indeed $2L\varepsilon$ transitions of the Metropolis-adjusted Langevin algorithm. In each case the resulting proposals will follow distinctly different distributions. Analysing the speed of different diffusion limits for the two methods on a reference class of models should give some insight here.

The marginal representation relates HMC more explicitly to the Metropolis-adjusted Langevin algorithm than has been reported previously, but also offers insight into the various tuning parameters of the method, not only the integration time but also the kinetic energy choice. It is possible that a different choice of kinetic energy than $p^t M^{-1} p/2$ could lead to different ergodicity properties, either more or less favourable, and this type of analysis is immediately feasible based on the framework introduced in this thesis.

Bibliography

[1] Shun-ichi Amari and Hiroshi Nagaoka. *Methods of information geometry*, volume 191. American Mathematical Soc., 2007.

[2] Christophe Andrieu and Éric Moulines. On the ergodicity properties of some adaptive MCMC algorithms. *The Annals of Applied Probability*, 16(3):1462–1505, 2006.

[3] Christophe Andrieu and Johannes Thoms. A tutorial on adaptive MCMC. *Statistics and Computing*, 18(4):343–373, 2008.

[4] Krishna B Athreya and P Ney. A new approach to the limit theory of recurrent Markov chains. *Transactions of the American Mathematical Society*, 245:493–501, 1978.

[5] OE Barndorff-Nielsen, DR Cox, and N Reid. The role of differential geometry in statistical theory. *International Statistical Review*, 54(1):83–96, 1986.

[6] Alexandros Beskos, Natesh Pillai, Gareth Roberts, Jesus-Maria Sanz-Serna, and Andrew Stuart. Optimal tuning of the hybrid Monte Carlo algorithm. *Bernoulli*, 19(5A):1501–1534, 2013.

[7] Alexandros Beskos, Gareth Roberts, Alexandre Thiery, and Natesh Pillai. Asymptotic Analysis of the Random-Walk Metropolis Algorithm on Ridged Densities. *arXiv preprint arXiv:1510.02577*, 2015.

[8] Michael Betancourt. A General Metric for Riemannian Manifold Hamiltonian Monte Carlo. In *Geometric Science of Information*, pages 327–334. Springer, 2013.

[9] MJ Betancourt. Generalizing the no-U-turn sampler to Riemannian manifolds. *arXiv preprint arXiv:1304.1920*, 2013.

[10] MJ Betancourt, Simon Byrne, and Mark Girolami. Optimizing the integrator step size for Hamiltonian Monte Carlo. *arXiv preprint arXiv:1411.6669*, 2014.

[11] MJ Betancourt, Simon Byrne, Samuel Livingstone, and Mark Girolami. The geometric foundations of Hamiltonian Monte Carlo. *arXiv preprint arXiv:1410.5110*, 2014.

[12] George D Birkhoff. Proof of the ergodic theorem. *Proceedings of the National Academy of Sciences*, 17(12):656–660, 1931.

[13] William M Boothby. *An introduction to differentiable manifolds and Riemannian geometry*, volume 120. Academic press, 1986.

[14] Nawaf Bou-Rabee, Aleksandar Donev, and Eric Vanden-Eijnden. Metropolis integration schemes for self-adjoint diffusions. *Multiscale Modeling & Simulation*, 12(2):781–831, 2014.

[15] Nawaf Bou-Rabee and Martin Hairer. Nonasymptotic mixing of the MALA algorithm. *IMA Journal of Numerical Analysis*, page drs003, 2012.

[16] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013.

[17] George EP Box and Mervin E Müller. A note on the generation of random normal deviates. *The annals of mathematical statistics*, 29(2):610–611, 1958.

[18] Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng. *Handbook of Markov Chain Monte Carlo*. CRC Press, 2011.

[19] Eric Cances, Frédéric Legoll, and Gabriel Stoltz. Theoretical and numerical comparison of some sampling methods for molecular dynamics. *ESAIM: Mathematical Modelling and Numerical Analysis*, 41(02):351–389, 2007.

[20] Marek Capinski and Peter E Kopp. *Measure, integral and probability*. Springer, 2004.

[21] Kung Sik Chan and Charles J Geyer. Discussion: Markov chains for exploring posterior distributions. *The Annals of Statistics*, pages 1747–1758, 1994.

[22] Ole F Christensen, Gareth O Roberts, and Martin Sköld. Robust Markov chain Monte Carlo methods for spatial generalized linear mixed models. *Journal of Computational and Graphical Statistics*, 15(1):1–17, 2006.

[23] William Coffey, Yu P Kalmykov, and John T Waldron. *The Langevin equation: with applications to stochastic problems in physics, chemistry, and electrical engineering*, volume 14. World Scientific, 2004.

[24] John D Cook. Upper and lower bounds on the Normal distribution function. Available at: `http://www.johndcook.com/normalbounds.pdf`. Accessed: 2015-06-29, October 2009.

[25] Radu V Craiu, Jeffrey Rosenthal, and Chao Yang. Learn from thy neighbor: Parallel-chain and regional adaptive MCMC. *Journal of the American Statistical Association*, 104(488):1454–1466, 2009.

[26] Frank Critchley, Paul Marriott, and Mark Salmon. Preferred point geometry and statistical manifolds. *The Annals of Statistics*, 21(3):1197–1224, 1993.

[27] A DasGupta. *Asymptotic Theory of Statistics and Probability*. Springer–Verlag New York, 2008.

[28] Persi Diaconis. The markov chain monte carlo revolution. *Bulletin of the American Mathematical Society*, 46(2):179–205, 2009.

[29] Persi Diaconis. Some things weve learned (about Markov chain Monte Carlo). *Bernoulli*, 19(4):1294–1305, 2013.

[30] Persi Diaconis, Susan Holmes, and Radford M Neal. Analysis of a nonreversible Markov chain sampler. *Annals of Applied Probability*, pages 726–752, 2000.

[31] Persi Diaconis, Christof Seiler, and Susan Holmes. Connections and Extensions: A Discussion of the Paper by Girolami and Byrne. *Scandinavian Journal of Statistics*, 41(1):3–7, 2014.

[32] Manfredo P Do Carmo. *Riemannian geometry*. Springer, 1992.

[33] Wolfang Doeblin. Exposé de la théorie des chaıne simples constantes de Markova un nombre fini détats. *Mathématique de lUnion Interbalkanique*, 2(77-105):78–80, 1938.

[34] Wolfgang Doeblin. Elements d'une theorie generale des chaines simples constantes de Markoff. In *Annales Scientifiques de l'Ecole Normale Supérieure*, volume 57, pages 61–111, 1940.

[35] Simon Duane, Anthony D Kennedy, Brian J Pendleton, and Duncan Roweth. Hybrid monte carlo. *Physics letters B*, 195(2):216–222, 1987.

[36] AB Duncan, T Lelievre, and GA Pavliotis. Variance Reduction using Nonreversible Langevin Samplers. *arXiv preprint arXiv:1506.04934*, 2015.

[37] Andreas Eberle. Error bounds for MetropolisHastings algorithms applied to perturbations of Gaussian measures in high dimensions. *Ann. Appl. Probab.*, 24(1):337–377, 02 2014.

[38] Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*, volume 57. CRC press, 1994.

[39] Gersende Fort and Eric Moulines. Polynomial ergodicity of Markov transition kernels. *Stochastic Processes and their Applications*, 103(1):57–99, 2003.

[40] Georg Ferdinand Frobenius. *Über Matrizen aus nicht negativen Elementen*. Königliche Akademie der Wissenschaften, 1912.

[41] Charles J Geyer. Markov chain Monte Carlo maximum likelihood. In *Computing Science and Statistics, Proceedings of the 23rd Symposium on the Interface*, pages 156–163. Interface Foundation of North America, 1991.

[42] Alison L Gibbs and Francis Edward Su. On choosing and bounding probability metrics. *International statistical review*, 70(3):419–435, 2002.

[43] Mark Girolami and Ben Calderhead. Riemann manifold langevin and hamiltonian monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214, 2011.

[44] Geoffrey Grimmett and David Stirzaker. *Probability and random processes*, volume 2. Oxford Univ Press, 1992.

[45] P Gudynas. Refinements of the central limit theorem for homogeneous Markov chains. In *Limit Theorems of Probability Theory*, pages 167–183. Springer, 2000.

[46] Heikki Haario, Eero Saksman, and Johanna Tamminen. An adaptive Metropolis algorithm. *Bernoulli*, pages 223–242, 2001.

[47] Theodore Edward Harris. The existence of stationary measures for certain Markov processes. In *Proceedings of the 3rd Berkeley Symposium on Mathematical Statistics and Probability*, volume 2, pages 113–124, 1956.

[48] W Keith Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[49] Matthew D Hoffman and Andrew Gelman. The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *The Journal of Machine Learning Research*, 15(1):1593–1623, 2014.

[50] Susan Holmes, Simon Rubinstein-Salzedo, and Christof Seiler. Curvature and concentration of Hamiltonian Monte Carlo in high dimensions. *arXiv preprint arXiv:1407.1114*, 2014.

[51] Elton P Hsu. *Stochastic analysis on manifolds*, volume 38. American Mathematical Soc., 2002.

[52] Søren F Jarner and Richard L Tweedie. Necessary conditions for geometric and polynomial ergodicity of random-walk-type. *Bernoulli*, 9(4):559–578, 2003.

[53] Harold Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 186(1007):453–461, 1946.

[54] Leif T Johnson and Charles J Geyer. Variable Transformation to Obtain Geometric Ergodicity in the Random-walk Metropolis Algorithm. *The Annals of Statistics*, 40(6):3050–3076, 2012.

[55] Norman L Johnson and Samuel Kotz. *Distributions in Statistics: Continuous Univariate Distributions: Vol.: 1*. Houghton Mifflin, 1970.

[56] Galin L Jones. On the Markov chain central limit theorem. *Probability surveys*, 1(299-320):5–1, 2004.

[57] Galin L Jones and James P Hobert. Honest exploration of intractable probability distributions via Markov chain Monte Carlo. *Statistical Science*, pages 312–334, 2001.

[58] Aldéric Joulin and Yann Ollivier. Curvature, concentration and error estimates for Markov chain Monte Carlo. *The Annals of Probability*, 38(6):2418–2442, 2010.

[59] Olav Kallenberg. *Foundations of modern probability*. Springer Science & Business Media, 2006.

[60] John Kent. Time-reversible diffusions. *Advances in Applied Probability*, 10(4):819–835, 1978.

[61] Andrei Nikolaevitch Kolmogorov. Markov chains with a countable number of possible states. *Byull. Mosk. Gos. Univ., Mat. Mekh*, 1(3):1–16, 1937.

[62] Krzystof Łatuszynski, Gareth O. Roberts, Alexandre Thiery, and Katarzyna Wolny. Discussion on 'Riemann manifold langevin and hamiltonian monte carlo methods' (by Girolami, M. and Calderhead, B.). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):188–189, 2011.

[63] Krzysztof Łatuszyński, Błażej Miasojedow, and Wojciech Niemiro. Nonasymptotic bounds on the estimation error of MCMC algorithms. *Bernoulli*, 19(5A):2033–2066, 2013.

[64] John M Lee. *Introduction to smooth manifolds*. Springer, 2003.

[65] Benedict Leimkuhler and Sebastian Reich. *Simulating hamiltonian dynamics*, volume 14. Cambridge University Press, 2004.

[66] David Asher Levin, Yuval Peres, and Elizabeth Lee Wilmer. *Markov chains and mixing times*. American Mathematical Soc., 2009.

[67] Thomas Milton Liggett. *Continuous time Markov processes: an introduction*, volume 113. American Mathematical Soc., 2010.

[68] Dennis V Lindley and Adrian FM Smith. Bayes estimates for the linear model. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–41, 1972.

[69] Torgny Lindvall. *Lectures on the coupling method*. Courier Corporation, 2002.

[70] Samuel Livingstone. Geometric ergodicity of the Random Walk Metropolis with position-dependent proposal covariance. *arXiv preprint arXiv:1507.05780*, 2015.

[71] Samuel Livingstone and Mark Girolami. Information-Geometric Markov Chain Monte Carlo Methods Using Diffusions. *Entropy*, 16(6):3074–3102, 2014.

[72] David JC MacKay. Introduction to monte carlo methods. In *Learning in graphical models*, pages 175–204. Springer, 1998.

[73] Jonathan H Manton. A Primer on Stochastic Differential Geometry for Signal Processing. *arXiv preprint arXiv:1302.0430*, 2013.

[74] Andrey Andreyevich Markov. Extension of the law of large numbers to dependent quantities. *Izv. Fiz.-Matem. Obsch. Kazan Univ.(2nd Ser)*, 15:135–156, 1906.

[75] Paul Marriott. On the local geometry of mixture models. *Biometrika*, 89(1):77–93, 2002.

[76] Jonathan C Mattingly, Andrew M Stuart, and Desmond J Higham. Ergodicity for SDEs and approximations: locally Lipschitz vector fields and degenerate noise. *Stochastic processes and their applications*, 101(2):185–232, 2002.

[77] Felipe J Medina-Aguayo, Anthony Lee, and Gareth O Roberts. Stability of Noisy Metropolis-Hastings. *arXiv preprint arXiv:1503.07066*, 2015.

[78] Kerrie L Mengersen and Richard L Tweedie. Rates of convergence of the Hastings and Metropolis algorithms. *The Annals of Statistics*, 24(1):101–121, 1996.

[79] Nicholas Metropolis and Stanislaw Ulam. The Monte Carlo method. *Journal of the American statistical association*, 44(247):335–341, 1949.

[80] Sean P Meyn and Richard L Tweedie. Stability of Markovian processes III: Foster-Lyapunov criteria for continuous-time processes. *Advances in Applied Probability*, 25:518–518, 1993.

[81] Sean P Meyn and Richard L Tweedie. *Markov Chains and Stochastic Stability*. Springer, 2008.

[82] Antonietta Mira. Ordering and improving the performance of Monte Carlo Markov chains. *Statistical Science*, pages 340–350, 2001.

[83] John F Nash Jr. The imbedding problem for Riemannian manifolds. *The Essential John Nash*, page 151, 2002.

[84] R Neal. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, pages 113–162, 2011.

[85] Radford M Neal. Slice sampling. *Annals of statistics*, pages 705–741, 2003.

[86] James R Norris. *Markov chains*. Cambridge university press, 1998.

[87] Esa Nummelin. A splitting technique for Harris recurrent Markov chains. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 43(4):309–318, 1978.

[88] Esa Nummelin. *General irreducible Markov chains and non-negative operators*, volume 83. Cambridge University Press, 2004.

[89] Bernt Øksendal. *Stochastic differential equations*. Springer, 2003.

[90] Yudi Pawitan. *In all likelihood: statistical modelling and inference using likelihood*. Oxford University Press, 2001.

[91] Oskar Perron. Zur theorie der matrices. *Mathematische Annalen*, 64(2):248–263, 1907.

[92] Peter H Peskun. Optimum monte-carlo sampling using markov chains. *Biometrika*, 60(3):607–612, 1973.

[93] Noemi Petra, James Martin, Georg Stadler, and Omar Ghattas. A computational framework for infinite-dimensional Bayesian inverse problems: Part II. Stochastic Newton MCMC with application to ice sheet flow inverse problems. *arXiv preprint arXiv:1308.6221*, 2013.

[94] C Radhakrishna Rao. Information and accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society*, 37(3):81–91, 1945.

[95] Jim O Ramsay, G Hooker, D Campbell, and J Cao. Parameter estimation for differential equations: a generalized smoothing approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(5):741–796, 2007.

[96] CE Rasmussen. Discussion of the paper Riemann manifold Langevin and Hamiltonian Monte Carlo methods by Mark Girolami and Ben Calderhead. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 73(2):161–162, 2011.

[97] Daniel Revuz. *Markov chains*. Elsevier, 2008.

[98] Brian D Ripley. *Stochastic simulation*, volume 316. John Wiley & Sons, 2009.

[99] Christian P Robert and George Casella. *Monte Carlo statistical methods*, volume 319. Citeseer, 2004.

[100] Gareth O Roberts. Linking theory and practice of MCMC. In *Highly structured stochastic systems*, pages 145–166. Oxford University Press, 2003.

[101] Gareth O Roberts, Andrew Gelman, and Walter R Gilks. Weak convergence and optimal scaling of random walk Metropolis algorithms. *The annals of applied probability*, 7(1):110–120, 1997.

[102] Gareth O Roberts and Jeffrey S Rosenthal. Geometric ergodicity and hybrid Markov chains. *Electron. Comm. Probab*, 2(2):13–25, 1997.

[103] Gareth O Roberts and Jeffrey S Rosenthal. On convergence rates of Gibbs samplers for uniform distributions. *Annals of Applied Probability*, pages 1291–1302, 1998.

[104] Gareth O Roberts and Jeffrey S Rosenthal. Optimal scaling for various Metropolis-Hastings algorithms. *Statistical science*, 16(4):351–367, 2001.

[105] Gareth O Roberts and Jeffrey S Rosenthal. General state space Markov chains and MCMC algorithms. *Probability Surveys*, 1:20–71, 2004.

[106] Gareth O Roberts and Jeffrey S Rosenthal. Examples of adaptive MCMC. *Journal of Computational and Graphical Statistics*, 18(2):349–367, 2009.

[107] Gareth O Roberts, Jeffrey S Rosenthal, et al. Minimising MCMC variance via diffusion limits, with an application to simulated tempering. *The Annals of Applied Probability*, 24(1):131–149, 2014.

[108] Gareth O Roberts and Osnat Stramer. Langevin diffusions and Metropolis-Hastings algorithms. *Methodology and computing in applied probability*, 4(4):337–357, 2002.

[109] Gareth O Roberts and Richard L Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, pages 341–363, 1996.

[110] Gareth O Roberts and Richard L Tweedie. Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika*, 83(1):95–110, 1996.

[111] L Chris G Rogers and David Williams. *Diffusions, Markov processes and martingales: Volume 2, Itô calculus*, volume 2. Cambridge university press, 2000.

[112] Walter Rudin. *Functional analysis, 1973*. McGraw-Hill, New York, 1973.

[113] Daniel Rudolf. Explicit error bounds for Markov chain Monte Carlo. *arXiv preprint arXiv:1108.3201*, 2011.

[114] Mark J Schervish. *Theory of statistics*. Springer, 1995.

[115] Bernard F. Schutz. *Geometrical methods of mathematical physics*. Cambridge University Press, 1984.

[116] Dino Sejdinovic, Heiko Strathmann, Maria Lomeli Garcia, Christophe Andrieu, and Arthur Gretton. Kernel adaptive Metropolis–Hastings. In *Proceedings of the 31st International Conference on Machine Learning*, 2014.

[117] Eugene Seneta. Markov and the birth of chain dependence theory. *International Statistical Review/Revue Internationale de Statistique*, pages 255–263, 1996.

[118] Eugene Seneta. Statistical regularity and free will: LAJ Quetelet and PA Nekrasov. *International statistical review*, 71(2):319–334, 2003.

[119] Lawrence F Shampine and Charles William Gear. A user's view of solving stiff ordinary differential equations. *SIAM review*, 21(1):1–17, 1979.

[120] Chris Sherlock. Optimal scaling of the random walk Metropolis: general criteria for the 0.234 acceptance rule. *Journal of Applied Probability*, 50(1):1–15, 2013.

[121] Chris Sherlock, Paul Fearnhead, and Gareth O Roberts. The random walk Metropolis: linking theory and practice through a case study. *Statistical Science*, 25(2):172–190, 2010.

[122] James Stewart. *Multivariable calculus*. Cengage Learning, 2011.

[123] Stan Development Team. *Stan Modeling Language Users Guide and Reference Manual, Version 2.7.0*, 2015.

[124] Luke Tierney. Markov chains for exploring posterior distributions. *the Annals of Statistics*, pages 1701–1728, 1994.

[125] Luke Tierney. A note on Metropolis-Hastings kernels for general state spaces. *Annals of Applied Probability*, pages 1–9, 1998.

[126] Lloyd N Trefethen and David Bau III. *Numerical linear algebra*, volume 50. Siam, 1997.

[127] Richard L Tweedie and Sean P Meyn. The Doeblin decomposition. In *Doeblin and Modern Probability*, volume 149, page 211. American Mathematical Soc., 1993.

[128] John von Neumann. Proof of the ergodic theorem and the H-theorem in quantum mechanics. *The European Physical Journal H*, 35(2):201–237, 2010.

[129] Eric W Weisstein. *"Eigenvector." From MathWorld–A Wolfram Web Resource*, 2015.

[130] Tatiana Xifara, Chris Sherlock, Samuel Livingstone, Simon Byrne, and Mark Girolami. Langevin diffusions and the Metropolis-adjusted Langevin algorithm. *Statistics and Probability Letters*, 91:14–19, 2014.

Appendix A

Some results on Markov chains.

We expand on some results stated in the main body.

Proof that " $\leftrightarrow$" defines an equivalence relation on a countable X.

We need to show (i) $x \leftrightarrow x$ for all states $x \in \mathbf{X}$, (ii) $x \leftrightarrow y \Rightarrow y \leftrightarrow x$, and (iii) if $x \leftrightarrow y$ and $y \leftrightarrow z$ then $x \leftrightarrow z$.

The first is true by the fact that $P^0(x,x) = 1$. The second is trivial since if $x \leftrightarrow y$ then $\exists m,n$ s.t. $P^m(x,y) > 0$ and $P^n(y,x) > 0$, and these properties are also the requirement for $y \leftrightarrow x$. The third can be seen using the Chapman-Kolmogorov equations: since there must exist $\exists r,s$ s.t. $P^r(x,y) > 0$ and $P^s(y,z) > 0$, it holds that $P^{r+s}(x,z) \geq P^r(x,y)P^s(y,z) > 0$, so that $x \to z$, and an equivalent argument shows $z \to x$. ■

Proof of Proposition 3.4

First note that

$$\mathbb{E}_x[\eta_x] = \mathbb{E}_x\left[\sum_{t=0}^{\infty} \mathbb{1}_x(X_t)\right] = \sum_{t=0}^{\infty} P^t(x,x),$$

where the last identity is true by monotone convergence (see e.g. Section 4.2. of [20]).

We show that $\sum_{t=0}^{\infty} P^t(x,x) = \infty \Rightarrow \mathbb{P}_x[\tau_x < \infty] = 1$. If we first consider the probability of a finite

occupancy time η_x, note that

$$1 \geq \mathbb{P}_x[\eta_x < \infty] = \sum_{t=0}^{\infty} \mathbb{P}_x[X_t = x \cap X_{t+i} \neq x, \ \forall i \geq 1].$$

Using the Markov property this becomes

$$1 \geq \sum_{t=0}^{\infty} \mathbb{P}[X_{t+i} \neq x, \ \forall i \geq 1 | X_t = x] P^t(x,x), \tag{A.1}$$

$$= \sum_{t=0}^{\infty} \mathbb{P}_x[\tau_x = \infty] P^t(x,x), \tag{A.2}$$

which will only be satisfied if $\mathbb{P}_x[\tau_x = \infty] = 0$, implying $\mathbb{P}_x[\tau_x < \infty] = 1$ as required. ■

The reverse implication can also be proved (see e.g. [86]).

Explicit construction of a maximal irreducibility measure.

We have a φ-irreducible chain for which we would like to find a maximal irreducibility measure $\psi(\cdot)$. We first define the resolvent transition kernel as

$$K_\varepsilon(x,A) = (1-\varepsilon) \sum_{i=1}^{\infty} \varepsilon^i P^i(x,A),$$

for any $A \in \mathscr{B}$ and $x \in \mathbf{X}$, for any fixed choice $\varepsilon < 1$. The resolvent captures information about the i step transition kernel for every choice of i. Now we can easily find a $\psi(\cdot)$ by computing

$$\psi(A) = \int K_\varepsilon(x,A) \varphi(dx),$$

for any choice of ε.

Proof that recurrence and transience are class properties for a countable X.

We show that if x is recurrent and $x \leftrightarrow y$ then y is recurrent. It follows that if x is transient then so is y, because if y were recurrent then x would be also.

Mathematically we want to establish that if $\mathbb{E}_x[\eta_x] = \infty$ and $x \leftrightarrow y$ then $\mathbb{E}_y[\eta_y] = \infty$ also. Since $x \leftrightarrow y$ then $\exists, k, m$ s.t. $P^k(x,y) > 0$ and $P^m(y,x) > 0$. By the Chapman-Kolmogorov equations

$$P^{k+m+n}(y,y) \geq P^m(y,x) P^n(x,x) P^k(x,y),$$

so summing across all possible n values gives

$$\mathbb{E}_y[\eta_y] \geq \sum_{n=0}^{\infty} P^{k+m+n}(y,y) \geq P^m(y,x) P^k(x,y) \mathbb{E}_x[\eta_x].$$

From this we have that y is recurrent if x is. ■

Proof of Theorem 3.22. We need to show $|\mu(A) - \nu(A)| \leq \mathbb{P}_\Lambda[X \neq Y]$, for any $A \in \mathscr{B}$. First note that $\mu(A) = \mathbb{P}_\mu[X \in A]$, and similarly for $\nu(A)$ and Y. Now

$$\mathbb{P}_\mu[X \in A] = \mathbb{P}_\Lambda[X \in A, X \neq Y] + \mathbb{P}_\Lambda[X \in A, X = Y],$$
$$\mathbb{P}_\nu[Y \in A] = \mathbb{P}_\Lambda[Y \in A, X \neq Y] + \mathbb{P}_\Lambda[Y \in A, X = Y],$$

and $\mathbb{P}_\Lambda[X \in A, X = Y] = \mathbb{P}_\Lambda[Y \in A, X = Y]$, so we can write

$$|\mu(A) - \nu(A)| = |\mathbb{P}_\Lambda[X \in A, X \neq Y] - \mathbb{P}_\Lambda[Y \in A, X \neq Y]|.$$

Now, for any $x, y \geq 0$ note that $|x - y| \leq \max\{x, y\}$, giving

$$|\mu(A) - \nu(A)| \leq \max\{\mathbb{P}_\Lambda[X \in A, X \neq Y], \mathbb{P}_\Lambda[Y \in A, X \neq Y]\}$$

Since $\mathbb{P}_\Lambda[X \in A, X \neq Y] = \mathbb{P}_\Lambda[X \neq Y]\mathbb{P}[X \in A | X \neq Y]$, and similarly for $\mathbb{P}_\Lambda[Y \in A, X \neq Y]$, we can write

$$|\mu(A) - \nu(A)| \leq \mathbb{P}_\Lambda[X \neq Y] \max\{\mathbb{P}[X \in A | X \neq Y], \mathbb{P}[Y \in A | X \neq Y]\} \leq \mathbb{P}_\Lambda[X \neq Y],$$

which completes the proof. ■.

Appendix B

Total variation distance

We show how to obtain (3.14) from (3.13). Denoting two probability distributions, $\mu(\cdot)$ and $\nu(\cdot)$, and associated densities, $\mu(x)$ and $\nu(x)$, we have

$$\|\mu(\cdot) - \nu(\cdot)\|_{TV} := \sup_{A \in \mathscr{B}} |\mu(A) - \nu(A)|.$$

Define the set $B = \{x \in \mathbf{X} : \mu(x) > \nu(x)\}$. To see that $B \in \mathscr{B}$, note that $B = \cup_{q \in \mathbb{Q}} \{x \in \mathbf{X} : \mu(x) > q\} \cap \{x \in \mathbf{X} : \nu(x) < q\}$, and the result follows from properties of $\mathscr{B}$ (see e.g. Section 2.5 of [20]). Now, for any $A \in \mathscr{B}$

$$\mu(A) - \nu(A) \leq \mu(A \cap B) - \nu(A \cap B) \leq \mu(B) - \nu(B),$$

and similarly

$$\nu(A) - \mu(A) \leq \nu(B^c) - \mu(B^c),$$

so, the supremum will be attained either at B or B^c. However, since $\mu(\mathbf{X}) = \nu(\mathbf{X}) = 1$, then

$$[\mu(B) - \nu(B)] - [\nu(B^c) - \mu(B^c)] = 0,$$

so that

$$|\mu(B) - \nu(B)| = |\mu(B^c) - \nu(B^c)|.$$

Using these facts gives an alternative characterisation of the total variation distance as

$$\begin{aligned} \|\mu(\cdot) - \nu(\cdot)\|_{TV} &= \frac{1}{2} \left(|\mu(B) - \nu(B)| + |\mu(B^c) - \nu(B^c)| \right) \\ &= \frac{1}{2} \int_{\mathbf{X}} |\mu(x) - \nu(x)| dx \end{aligned}$$

as required.

Appendix C

Some objects from Riemannian geometry

We provide more details on some generalisations of objects in $\mathbb{R}^n$ to Riemannian manifolds.

Gradient and divergence operators

The gradient of a function on $\mathbb{R}^n$ is the unique vector field, such that, for any unit vector, u:

$$\langle \nabla f(x), u \rangle = D_u [f(x)] = \lim_{h \to 0} \left\{ \frac{f(x+hu) - f(x)}{h} \right\}, \tag{C.1}$$

the directional derivative of f along u at $x \in \mathbb{R}^n$.

On a manifold, the gradient operator, ∇_M, can still be defined, such that the inner product $g_p(\nabla_M f(x), u) = D_u[f(x)]$. Setting $\nabla_M = G(x)^{-1}\nabla$ gives:

$$\begin{aligned} g_p(\nabla_M f(x), u) &= (G^{-1}(x)\nabla f(x))^T G(x) u, \\ &= \langle \nabla f(x), u \rangle, \end{aligned}$$

which is equal to the directional derivative along u as required.

The divergence of some vector field, v, at a point, $x \in \mathbb{R}^n$, is the net outward flow generated by v through some small neighbourhood of x. Mathematically, the divergence of $v(x) \in \mathbb{R}^3$ is given by $\sum_i \partial v_i / \partial x_i$. On a more general manifold, the divergence is also a sum of derivatives, but here, they

are covariant derivatives. A short introduction is provided in Appendix C. Here, we simply state that the covariant derivative of a vector field, v, at a point $p \in M$ is the orthogonal projection of the directional derivative onto the tangent space, T_pM. Intuitively, a vector field on a manifold is a field of vectors, each of which lie in the tangent space to a point, $p \in M$. It only makes sense therefore to discuss how vector fields change along the manifold or in the direction of vectors, which also lie in the tangent space. Although the idea seems simple, the covariant derivative has some attractive geometric properties; notably, it can be completely written in local coordinates,and, so, does not depend on knowledge of an embedding in some ambient space.

The divergence of a vector field, v, defined on a manifold, M, at the point, $p \in M$, is defined as:

$$\mathrm{div}_M(v) = \sum_{i=1}^{n} D^c_{e_i}[v_i],$$

where e_i denotes the i-th basis vector for the tangent space, T_pM, at $p \in M$, and v_i denotes the i-th coefficient. This can be written in local coordinates (see Appendix C) as:

$$\mathrm{div}_M(v) = |G(x)|^{-\frac{1}{2}} \sum_{i=1}^{n} \frac{\partial}{\partial x_i} \left(|G(x)|^{\frac{1}{2}} v_i \right),$$

and can be combined with ∇_M to form the Laplace–Beltrami operator (5.8).

Vector fields and the covariant derivative

Here, we provide a short introduction to vector fields and differentiation on a smooth manifold; see [13, 64]. The following geometric notation is used here: (i) vector components are indexed with a superscript, e.g., $v = (v^1, ..., v^n)$; and (ii) repeated subscripts and superscripts are summed over, e.g., $v^i \mathbf{e}_i = \sum_i v^i \mathbf{e}_i$ (known as the Einstein summation convention).

For any smooth manifold, M, the set of all tangent vectors to points on M is known as the tangent bundle and denoted TM.

A C^r vector field defined on M is a mapping that assigns to each point, $p \in M$, a tangent vector, $v(p) \in T_pM$. In addition, the components of $v(p)$ in any basis for T_pM must also be C^r [13]. We will denote the set of all vector fields on M as $\Gamma(TM)$. For some vector field, $v \in \Gamma(TM)$, at any point, $p \in M$, the vector, $v(p) \in T_pM$, can be written as a linear combination of some n basis vectors $\{\mathbf{e}_1, ..., \mathbf{e}_n\}$ as $v = v^i \mathbf{e}_i$. To understand how v will change in a particular direction along M, it only makes sense, therefore, to consider derivatives along vectors in T_pM. Two other things must be considered when defining a derivative along a manifold: (i) how the components, v^i, of each basis

vector will change; and (ii) how each basis vector, $\mathbf{e}_i$, itself will change. For the usual directional derivative on $\mathbb{R}^n$, the basis vectors do not change, as the tangent space is the same at each point, but for a more general manifold, this is no longer the case: the $\mathbf{e}_i$'s are referred to as a "local" basis for each T_pM.

The covariant derivative, D^c, is defined so as to account for these shortcomings. When considering differentiation along a vector, $u^* \notin T_pM$, u^* is simply projected onto the tangent space. The derivative with respect to any $u \in T_pM$ can now be decomposed into a linear combination of derivatives of basis vectors and vector components:

$$D^c_u[v] = D^c_{u^i\mathbf{e}_i}[v^i\mathbf{e}_i], \tag{C.2}$$

where the argument, p, has been dropped, but is implied for both components and local basis vectors. The operator, $D^c_u[v]$, is defined to be linear in both u and v and to satisfy the product rule [13]; so, Equation (C.2) can be decomposed into:

$$D^c_u[v] = u^i\left(D^c_{\mathbf{e}_i}[v^j]\mathbf{e}_j + v^j D^c_{\mathbf{e}_i}[\mathbf{e}_j]\right). \tag{C.3}$$

The operator, D^c, need, therefore, only be defined along the direction of basis vectors $\mathbf{e}_i$ and for vector component v^i and basis vector $\mathbf{e}_i$ arguments.

For components v^i, $D^c_{\mathbf{e}_j}[v^i]$ is defined as simply the partial derivative $\partial_j v^i := \partial v^i/\partial x^j$. The directional derivative of some basis vector $\mathbf{e}_i$ along some $\mathbf{e}_j$ is best understood through the example of a regular surface $\Sigma \subset \mathbb{R}^3$. Here, $D_{\mathbf{e}_j}[\mathbf{e}_i]$ will be a vector, $w \in \mathbb{R}^3$. Taking the basis for this space at the point, p, as $\{\mathbf{e}_1, \mathbf{e}_2, \hat{\mathbf{n}}\}$, where $\hat{\mathbf{n}}$ denotes the unit normal to $T_p\Sigma$, we can write $w = \alpha\mathbf{e}_1 + \beta\mathbf{e}_2 + \kappa\hat{\mathbf{n}}$. The covariant derivative, $D^c_{\mathbf{e}_j}[\mathbf{e}_i]$, is simply the projection of w onto $T_p\Sigma$, given by $w^* = \alpha\mathbf{e}_1 + \beta\mathbf{e}_2$. More generally, at some point, p, in a smooth manifold, M, the covariant derivative $D^c_{\mathbf{e}_j}[\mathbf{e}_i] = \Gamma^k_{ji}\mathbf{e}_k$ (with upper and lower indices summed over). The coefficients, Γ^k_{ji}, are known as the Christoffel symbols: Γ^k_{ji} denotes the coefficient of the k-th basis vector when taking the derivative of the i-th with respect to the j-th. If a Riemannian metric, g, is chosen for M; then, they can be expressed completely as a function of g (or in local coordinates as a function of the matrix, G). Using these definitions, Equation (C.3) can be re-written as:

$$D^c_u[v] = u^i\left(\partial_i v^k + v^j\Gamma^k_{ij}\right)\mathbf{e}_k. \tag{C.4}$$

The divergence of a vector field, $v \in \Gamma(TM)$, at the point, $p \in M$, is given by:

$$\mathrm{div}_M(v) = D^c_{\mathbf{e}_i}[v^i], \tag{C.5}$$

where, again, repeated indices are summed over. If $M = \mathbb{R}^n$, this reduces to the usual sum of partial derivatives, $\partial_i v^i$. On a more general manifold, M, the equivalent expression is:"'

$$D^c_{\mathbf{e}_i}[v^i] = \partial_i v^i + v^i \Gamma^j_{ij}, \tag{C.6}$$

where, again, repeated indices are summed. As has been previously stated, if a metric, g, and coordinate chart is chosen for M, the Christoffel symbols can be written in terms of the matrix, $G(x)$. In this case [115]:

$$\Gamma^j_{ij} = |G(x)|^{-\frac{1}{2}} \partial_i \left(|G(x)|^{\frac{1}{2}} \right), \tag{C.7}$$

so Equation (C.6) becomes:

$$D^c_{\mathbf{e}_i}[v^i] = |G(x)|^{-\frac{1}{2}} \partial_i \left(|G(x)|^{\frac{1}{2}} v^i \right), \tag{C.8}$$

where $v = v(x)$.

Appendix D

Needed facts about truncated Gaussian distributions

Here we collect some elementary facts used in the main text. For more detail see e.g. [55]. If X follows a truncated Gaussian distribution $N^T_{[a,b]}(\mu,\sigma^2)$ then it has density

$$f(x) = \frac{1}{\sigma Z_{a,b}}\phi\left(\frac{x-\mu}{\sigma}\right)\mathbb{1}_{[a,b]}(x),$$

where $\phi(x) = e^{-x^2/2}/\sqrt{2\pi}$, $\Phi(x) = \int_{-\infty}^{x}\phi(y)dy$ and $Z_{a,b} = \Phi((b-\mu)/\sigma) - \Phi((a-\mu)/\sigma)$. Defining $B = (b-\mu)/\sigma$ and $A = (a-\mu)/\sigma$, we have

$$\mathbb{E}[X] = \mu + \frac{\phi(A)-\phi(B)}{Z_{a,b}}\sigma$$

and

$$\mathbb{E}[e^{tX}] = e^{\mu t+\sigma^2t^2/2}\left[\frac{\Phi(B-\sigma t)-\Phi(A-\sigma t)}{Z_{a,b}}\right].$$

In the special case $b=\infty$, $a=0$ this becomes $e^{\mu t+\sigma^2t^2/2}\Phi(\sigma t)/Z_{a,b}$.

Appendix E

A simple bound on the Normal distribution function

This is reproduced here from [24] for ease of exposition. Consider the complementary cumulative distribution function for $Z \sim N(0,1)$, given by

$$\Phi^c(z) = \mathbb{P}[Z > z] = \frac{1}{\sqrt{2\pi}} \int_z^\infty e^{-t^2/2} dt.$$

An upper bound for $z \geq 0$ can be derived as

$$\sqrt{2\pi}\Phi^c(z) = \int_z^\infty e^{-t^2/2} dt < \int_z^\infty \frac{t}{z} e^{-t^2/2} dt = \frac{1}{z} e^{-z^2/2}.$$

For the lower bound, define

$$g(z) = \Phi^c(z) - \frac{1}{\sqrt{2\pi}} \frac{z}{z^2+1} e^{-z^2/2}.$$

We show that $g(z) > 0$ for $z \geq 0$. First note that $g(0) = 0.5 > 0$, and $\lim_{z\to\infty} g(z) = 0$. The derivative of g is

$$g'(z) = -\frac{2}{(z^2+1)^2} e^{-z^2/2} < 0.$$

Since the derivative is strictly decreasing on $[0,\infty)$, this gives a lower bound

$$\sqrt{2\pi}\Phi^c(z) \geq \frac{z}{z^2+1} e^{-z^2/2},$$

as required.